BBRC

Bioscience Biotechnology Research Communications

Special Issue Vol 13 Number (11) 2020
Print ISSN: 0974-6455
Online ISSN: 2321-4007
CODEN BBRCBA
www.bbrc.in
University Grants Commission (UGC)
New Delhi, India Approved Journal

Bioscience Biotechnology Research Communications
Special Issue Volume 13 Number (11) 2020

Special Issue Volume 13 Number (11) 2020

On

Microscopy and Image Analysis for Computational and Agricultural Biotechnology

An International Peer Reviewed Open Access Journal

Published By:
Society For Science and Nature
Bhopal, Post Box 78, GPO,
462001 India

Indexed by Thomson Reuters, Now Clarivate Analytics USA

SJIF 2020=7.728
Online Content Available:
Every 3 Months at www.bbrc.in

SOCIETY FOR SCIENCE AND NATURE
SCIENCE FOR LIFE

Registered with the Registrar of Newspapers for India under Reg. No. 498/2007
Bioscience Biotechnology Research Communications
Special Issue Volume 13 No (11) 2020

ISSN 0974-6455

Registered with the Registrar of Newspapers for India under Reg. No. 498/2007
Bioscience Biotechnology Research Communications
Special Issue Volume 13 No (11) 2020

EDITORIAL COMMUNICATION

The development of fluorescent probes and application of high-resolution optical microscopes biological image processing techniques became more reliable with a profound impact on research in the biological sciences. In addition, those imaging tools plays crucial role in understanding various sub-cellular characteristics that lead to development of the field of cell biology. This special issue is designed in the manner to understand the basics of the fluorescence microscopy, in particular imaging and analysing the subcellular machineries of various cell types. Furthermore, Infectious diseases are caused by wide range of pathogenic bacterial strains. Bacterial strains such as *Enterococcus faecalis* and *Pseudomonas aeruginosa* are closely associated with Urinary Tract Infection (UTI) and Cystic Fibrosis (CF- a chronic lung infection), respectively. Therefore, this special issue aims to provide insights to the participants on recent trends in identification of bacterial cells and its subcellular proteins using modern techniques and below are some of non-exahustive list of techniques covered in this special issue.

- Fluorescence imaging
- F techniques
- Live cell tracking
- High throughput image analysis
- Artificial Intelligence and healthcare analytics
- Predictive analytics of diseases
- Healthcare security
- IoT in healthcare

Guest Editors

Dr. P. Saravanan Assistant Professor
of Biotechnology Rajalakshmi Engineering College Rajalakshmi Nagar Thandalam
Chennai 602 105, India
Email: saravanan.p@rajalakshmi.edu.in

Dr. B A Gowri Shankar Assistant Professor
School of Chemical & Biotechnology SASTRA Deemed University Tirumalaisamudram Thanjavur
Email: gowrishankar@scbt.sastra.edu

Dr.N.Rajathi Professor,
Department of Information Technology
Kumaraguru College of Technology Coimbatore
Chinnavedampatti, Coimbatore - 641049,
Tamil Nadu.Mobile:+919600558348
Email: rajathi.n.it@kct.ac.in

Dr.K.Kumaravel Assistant Professor,
Department of BioTechnology
Kumaraguru College of Technology Coimbatore
Chinnavedampatti, Coimbatore - 641049,
Tamil Nadu. Mobile: +917708257489
Email: kumaravel.k.bt@kct.ac.in

CONTENTS

BBRC
Bioscience Biotechnology Research Communications

VOLUME 13 • NUMBER (11) • SPECIAL ISSUE 2020

Biosc.Biotech.Res.Comm. Special Issue Vol 13 No 11 (2020) Pp-01-04

BBRC
Bioscience Biotechnology Research Communications

A Deep Learning Classifier for Accurate Detection of the Novel Corona Virus

Cynthia Jayapal[1]*, M. Sathiya Prakash[2] and M. Shiddharth Saran[3]
[1]*Professor, Department of Computer Science and Engineering,*
[2,3]*UG Scholars, Department of Computer Science and Engineering,*
Kumaraguru College of Technology, Coimbatore, India

ABSTRACT

Although India witnessed the second slowest 100 to 1000 jump in COVID-19 cases, according to WHO, the number may be inaccurate because of the lack of rapid and large-scale testing facilities. According to reports, India is yet to face the gruesome effects of this pandemic as it moves closer to stage 4 of the community spread. Though standardized tests used in detecting coronaviruses, such as RT- PCR or transcriptase-polymerase chain reaction, take a minimum of 24 hours to generate useful results, they are also prone to high false negatives. Consequently, multiple periodic tests are required to arrive at a firm confirmation. Owing to this gap in the Indian Coronavirus testing scenario, this study focuses on a comparatively rapid and accurate method of testing employing AI-based image analysis of X-Ray and CT scans of the Lungs. Artificial intelligence based deep learning methodologies involving Convolutional Neural Networks with a sharp eye on accuracy of results and practical usage could be used for image analysis. Pre-trained and well-known convolutional neural networks along with a standard dataset for training and testing the same have been selected for the process. The performance of the model is also analyzed using standardized convolutional neural network analysis techniques to infer the best model for the particular use-case. The main objective of the study is to evaluate whether deep learning has the potential to provide accurate results and could provide aid to the existing X-ray methodology.

KEY WORDS: CONVOLUTIONAL NEURAL NETWORKS, COVID-19, DEEP LEARNING, IMAGE ANALYSIS.

INTRODUCTION

Coronavirus (COVID-19) has produced rampant unprecedented decimation with millions of people losing their lives (Sohrabi, C. et al., 2020). Being a member of the SARS family of viruses, which surfaced in the early 2000s, the COVID - 19 strain has been particularly deadly. Claimed to be originated in Wuhan (China), this strain has affected 214 countries and territories around the world along with two international conveyances. As per a publication by the infamous John Hopkins University, under extreme conditions, up to 10% of the Indian population could be susceptible to getting affected by COVID-19. This puts things into perspective as to why India is in dire need to contain the pandemic before it reaches stage 4.

Thus, to achieve the same, we need rapid and accurate testing methods, contrary to the current time-intensive and inefficient methods (Yang, W. Yan, F. Patients et al.,2020) (Fang, Y. et al., 2020) (Xie, X. et al.,2020) to identify and isolate potential cases. From identifying the Zeroth patient in January to the current status(September 10, 2020), India has seen a 6000-fold rise in COVID-19 cases and they still continue to grow linearly. This number has been reached by testing only through methods such as swab tests and blood tests. With a population of

ARTICLE INFORMATION

**Corresponding Author:* cynthia.j.it@kct.ac.in
Received 10th Oct 2020 Accepted after revision 25th Nov 2020
Print ISSN: 0974-6455 Online ISSN: 2321-4007 CODEN: BBRCBA

Thomson Reuters ISI Web of Science Clarivate Analytics USA and Crossref Indexed Journal

NAAS Journal Score 2020 (4.31) SJIF: 2020 (7.728)

Online Contents Available at: http//www.bbrc.in/
Doi: http://dx.doi.org/10.21786/bbrc/13.11/1

135 crores, studies suggest that up to 10% of India's population (13.5 crores) are prone to COVID-19. The key is to test as many potential cases as possible. The present test which is done for the population rate is 18:1000000 which is too low because of the lack of kits.

Radiologists who deal with the diagnosis of the disease have observed common characteristics of physical conditions in the affected patients. The prominent method that is used to scan the affected body such as fractures, bone dislocations, lung infections, pneumonia, and tumors are X-ray machines. A computerized tomography (CT) scan (Yu, M. et al.,2020) combines a series of X-ray images taken from different angles around the body and uses computer processing to create cross-sectional images (slices) of the bones, blood vessels and soft tissues inside your body. Using X-rays (Gomez P et al.,2019) is a faster, easier, cheaper, and less harmful method compared to CT scans. The proposed solution is implemented by classifying based on X-rays images. Since X-ray machines and diagnostic centers are in comfortably higher numbers and can test at the cost of less than INR 400 per test, also being reusable, unlike the existing testing methodologies. Thus the serviceable addressable market for this solution would be the people who have limited access to present testing methods and have the symptoms of COVID-19.

This could be employed to perform periodic tests for frontline workers involved in this fight against the pandemic and post check-up for COVID-19 survivors. Keeping in mind the present state of the outbreak and crisis involved, saving lives is more important though the solution has high monetizability. This method could also serve as a preliminary check for potential COVID-19 patients before performing the actual swab/blood test which is limited in number. Thus we can save more PCR testing kits for more potential and vulnerable high risk patients. This method could highly improve the throughput of the patients checked everyday and thus, help us in isolating them and containing the outbreak. Therefore, this study aims at providing assistance to the existing X-ray methodology (Huang C et al.,2020) by proposing an accurate detection system for effective diagnosis in these tested times.

Since the problem scope addresses a classification problem(classifying people as either COVID-19 affected or not), the prominent approach to solving this could be done using classic AI. Artificial Intelligence (AI) (Negassi M et al.,2020) based automated CT and X-Ray image analysis tools will be of prominent use in the detection, quantification, and monitoring of the novel coronavirus. This method will effectively aid in distinguishing patients with coronavirus from benign ones. Based on the COVID-19 radiographic changes from CT images, advanced deep learning methods (Ching T et al.,2020) could extract the graphical features of COVID-19 induced changes in the Lungs. This provides clinical diagnosis prior to pathogenic testing and thus saves critical time in disease diagnosis. Deep learning is a subset of artificial intelligence that uses algorithmic structures called neural networks to perform classification tasks.

MATERIAL AND METHODS

Owing to its effective tunnel-shaped approach to network building, Convolutional neural networks are a patent choice for image classification. For this study, a particular variation of CNN based on pre-trained models(Huang, G. et al.,2016) that include ResNet50 ,VGG-16, and VGG-19 was trained on a dataset consisting of chest X-ray radiographs. Using this classifier, hospitals could classify the severity and prevent many COVID-19 fatalities by preparing earlier by doing whatever they can to stop the virus before their symptoms get too serious. A total of 400 CT scan images were collected from the LIDC open-source public dataset. (Armato et al.,2011) The functional steps of the classifier are as follows:

- When the CNN based classifier takes an X-ray radiograph as input, the slices are first passed through a feature extractor based on a preprocessed ResNet50, VGG16, and VGG19. Given the slices per image, the result is a tensor of shape (s, 256, 7, 7) for all the preprocessed models. Note: (256, 7, 7) is simply a set of 256 features maps of size 7x7 obtained by the last convolutional layer of the model.
- Each tensor of shape, for example- (256,7, 7), is then reduced in a vector of sh (256,) by computing the mean value of each 7x7 square by global average pooling. The shape of the output is (s, 256).
- In order to turn the (s, 256)-shaped tensor into a column vector before passing it to the classification part, max pooling is applied across the slices. A vector of shape (256,). is obtained.
- A column vector that is obtained is passed through fully connected layers with Softmax activation functions and dropout.
- The exact form of a datapoint varies between tasks, it could be a single image, a slice of a time series, a tabular record, or something else entirely. These are then passed on to a data loader that handles batching of data points and parallelism.
- A series of geometric transformations are carried out on each input of the radiographic image.
- The transformations are label-invariant and are meant to bring diversity to the dataset thereby increasing the stability of the model while decreasing its tendency to overfitting.
- Three geometric transformations are sequentially applied on each input image.
- Random rotation between -25 and 25 degrees is

applied.

- The random shift in both direction between -25 and 25 pixels
- Random horizontal flip with 50 percent probability
- The data obtained is then used to incrementally improve the model's ability to further classify the images.
- Random weights and biases are assigned during training the model. The results come out pretty poorly. The weights and biases are subsequently adjusted so as to obtain more correct predictions.
- The training of the model is done through the minimization of the cross-entropy loss using Adam optimizer.

RESULTS AND DISCUSSION

Figure 1: VGG -16 parameters

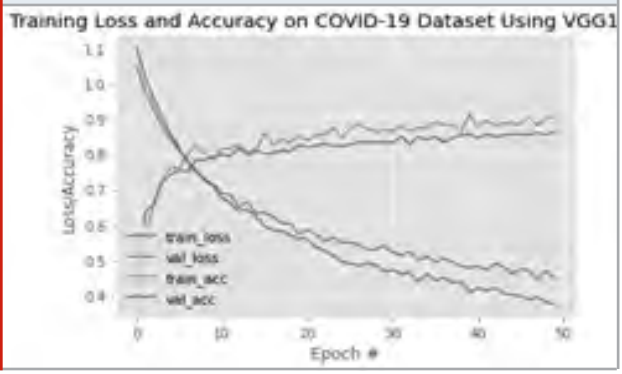


Figure 2: VGG -19 parameters

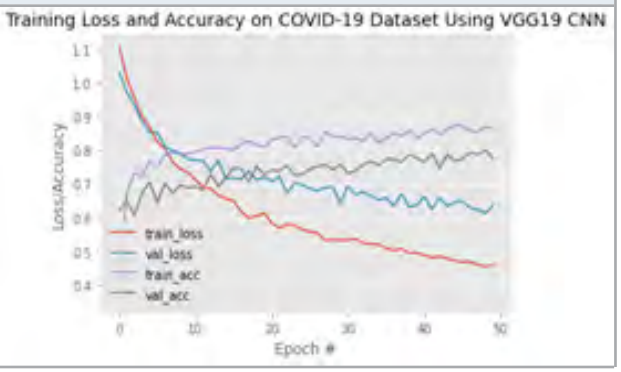
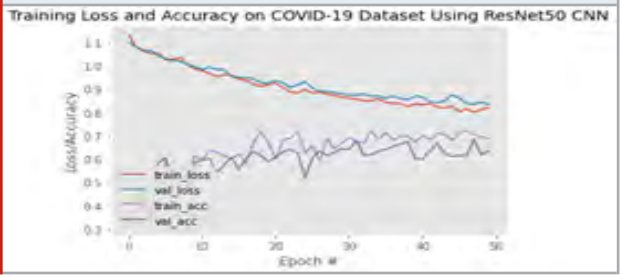


Figure 3: ResNet50 parameters



The models were subjected to Precision-Recall and confusion matrices analysis. The following results were inferred. (Refer Table 1)

The classifier was built with the best image classification deep neural networks with the most valid parameters to test them against. It can also be inferred that the chest X-ray images are better suited than the current methods for the detection of novel coronavirus. It is also observed that the pre-trained VGG16 model (Refer Figure 1) provides a high classification performance with a validation accuracy of 87% followed by VGG-19 (Refer Figure 2) and ResNet50 (Refer Figure 3)

As the COVID-19 cases keep rising, our country is nowhere near flattening the curve. Through this proposed system, as a supplement to the existing X-ray methodology, we could improve the diagnostic accuracy, specificity, and also reduce the diagnostic time required for the clinical experts. With the proposed system's accuracy, it could effectively replace the current time-intensive methods thereby better equipping India in this fight against COVID-19.

Table 1. Classification Analysis

MODEL		Precision	Recall	F1-score
VGG -16	Covid-19	0.35	0.34	0.35
	Normal	0.38	0.52	0.44
VGG-19	Covid-19	0.31	0.22	0.25
	Normal	0.38	0.52	0.44
ResNet50	Covid-19	0.33	0.49	0.39
	Normal	0.30	0.41	0.35

REFERENCES

Armato, S. G. 3rd et al. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. Med. Phys. 38, 915–931 (2011)

Ching T, Himmelstein DS, Beaulieu- Jones BK, Kalinin AA, Do BT, Way GP, et al. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface. 2018;15(141):20170387. doi:10.1098/rsif.2017.0387 (2018)

Fang, Y. et al. Sensitivity of chest CT for COVID-19: comparison to RT-PCR. Radiology. https://doi.org/10.1148/ radiol.2020200432 (2020)

Gomez P, Semmler M, Schutzenberger A, Bohr C, Dollinger M. Low-light image enhancement of high-speed endoscopic videos using a convolutional neural network. Med Biol Eng Comput. (2019)

Huang, G., Liu, Z., van der Maaten, L. Weinberger, K. Q. Densely connected convolutional networks (2016).

Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet. (2020)

Negassi M, Suarez-Ibarrola R, Hein S, Miernik A, Reiterer A. Application of artificial neural networks for automated analysis of cystoscopic images: a review of the current status and future prospects. World J Urol. (2020)

Sohrabi, C. et al. World Health Organiza- ion declares global emergency: a review of the 2019 novel coronavirus (COVID-19). Int J. Surg. 76, 71–76 (2020)

Xie, X. et al. Chest CT for typical 2019-nCoV pneumonia:relationship to negative RT-PCR testing. Radiology. https://doi.org/10.1148/ radiol.2020200343 (2020).

Yang, W. Yan, F. Patients with RT- PCR-confirmed COVID-19 and normal chest CT. Radiology 295, E3 (2020).

Yu, M. et al. Thin-section chest CT imaging of coronavirus disease 2019 pneumonia: comparison between patients with mild and severe disease. Radiol. Cardiothorac. Imaging 2 https:// doi.org/10.1148/ ryct.2020200126 (2020).

Biosc.Biotech.Res.Comm. Special Issue Vol 13 No 11 (2020) Pp-05-09

BBRC
Bioscience Biotechnology Research Communications

Agri Image Processing using Uml Model

Suguna M[1], S. Nithya Priya[2] and G. Thenmozhi[3]
[1]*Associate Professor, Information Science and Engineering Department, Kumaraguru College of Technology, Coimbatore, India*
[2]*Assistant Professor (SrG), Biotechnology Department, Kumaraguru College of Technology, Coimbatore, India*
[3]*Associate Professor, Automobile Engineering Department, Kumaraguru College of Technology, Coimbatore, India*

ABSTRACT

Agri image processing is an undertaking which points in building up a mechanized framework to separate and investigate leafy foods as great and terrible through an image handling framework. To check the nature of Agri-products through image processing handling utilizing MATLAB ide. It points in lessening human endeavors in the field of farming where there is an enormous region of land and it cannot check every yield physically, in this way fruits quality to check the nature of the product. Generally, this task of own is being created to support the customers, Businesspeople and Industrialists spare their time. It additionally helps in conveying great quality items to the Clients. This venture is helpful for both the merchants and purchasers of Agri-items.

KEY WORDS: AGRI PRODUCT, FRUIT ANALYSIS, FRUIT, SORTING, GRADING, PSNR.

INTRODUCTION

Agri image handling is a venture which points in building up a mechanized framework to separate and examine leafy foods as great and find the quality of the product through an image processing Preparing Framework. In this cycle to recognize the Nature of Horticulture items whether they are Positive or negative by utilizing the IAQ (image quality evaluation) using various parameters of pixel composition ratio Filter and measure MSE, PSNR, SNR of single image. In image methoding, helpless picture quality is pitiful for viable component extraction, include examination, design acknowledgment and quantitative mensuration. the photos are unexceptionally defiled by irregular commotion that occurs all through the estimation cycle in this manner confusing the robotized highlight extraction and examination of clinical Therefore, noisy images disposal might be an ought to for clinical pictures cycle to dispose of such demands while retentive the greatest sum as likely the vital picture[10]. Test cases are used to find the product good quality or bad quality.

MATERIAL AND METHODS

Literature Survey: The investigation done by a few scientists in the zone of picture order, foods grown from the ground characterization, natural products acknowledgment, natural product sickness [2] have proposed framework which discovers size of various products of the soil various natural products can be arranged utilizing fluffy rationale, here creator proposed MATLAB for the highlights extraction and for creating product of the soil order and natural product upset ID is viewed as a happening of image arrangement. The overwhelming majority of the explorers within the field of natural product acknowledgment or natural product

ARTICLE INFORMATION

*Corresponding Author: suguna.m.cse@kct.ac.in
Received 12th Oct 2020 Accepted after revision 24th Nov 2020
Print ISSN: 0974-6455 Online ISSN: 2321-4007 CODEN: BBRCBA

Thomson Reuters ISI Web of Science Clarivate Analytics USA and Crossref Indexed Journal

NAAS Journal Score 2020 (4.31) SJIF: 2020 (7.728)

*Online Contents Available at: http//www.bbrc.in/
Doi: http://dx.doi.org/10.21786/bbrc/13.11/2*

illness location have thought-about tone and surface properties for the order [5]. Review and management quality, one should have the option to quantify quality-related credits. During this paper [7,8] because of the high damp content in the product of the soil, water rules X-beam ingestion. A person's critical examination and reviewing of manufacture is usually a piece escalated, dreary, tedious, and emotional enterprise [6]. nevertheless, its costs, this system is variable, and decisions are not typically inevitable between assessors or from everyday [4]. The laptop vision examination of food things to be sensible, skilled, and reliable.

Proposed System: Projected framework explanation for this report is to assemble an application to visualize nature of Agri-items through image handling utilizing MATLAB ide. As huge sections of land cannot be checked physically, subsequently picture handling is utilized to check the nature of the yield. The accompanying strides to construct the following three steps. Step-1: A landing page which shows different alternatives (about, help, cycle, exit.) and screen where you can transfer a picture of the yield.

Step-2: Then convert the offered picture to highly contrasting for preparing and office to change over a given picture into good size and goal at that point apply a channel for extra handling of data. Step-3: Analysis of the picture depends on a couple of highlights like MSE, PSNR, SNR, SC... and so on and screen which shows the eventual outcome of the picture as fortunate or unfortunate.

Table. 2.1 Requirement Engineering

Functional requirements	Non-functional requirements
Product Perspective A distributed image processing database system includes description of the product and stores the following information.	Performance Perspective To store all the farming information base are as recorded underneath UML diagram
Product Features Agricultural yields are tested by traditional methods which consume labour. On the other hand, image processing techniques help to check quality of yields thus reducing manual labour.	Safety Requirements If there is broad harm to a wide segment of the information base because of disastrous disappointment.
User Class: Clients of the framework ought to have the option to recover yield data between given agricultural- products with the given image related data from the database. The software can be used by both sellers and buyers.	Security Requirements Security frameworks need information base stockpiling simply like numerous different applications. In any case, the exceptional prerequisites of the security market imply that merchants must pick their information base accomplice cautiously.
Working Environment Working climate for the quality testing framework is as recorded using the MATLAB platform. Design and Implementation Constraints	Software Quality Attributes Correctness: The quality of the processed image must be accurate. Maintainability and Usability: All the data must be stored safely, and the product ought to fulfil the greatest number of client's needs. It ought to be helpful to utilize.

Preprocessing: Pictures gained by different kinds of methods comprise of numerous commotions which disintegrate the part of a picture. In this manner, it cannot contribute fitting information for picture preparing. The Preprocessing upgrades the picture information, which beat hesitant twists and grow the highlights of the picture that are fundamental for handling and manufacturing an important picture than the first for a clear application. Component pre-processing modifications over associate degree information picture into a yield picture with the top goal that every yield pixel is said to the data pixel having the examination The farthest regular technique for pixel pre-preparing is shading for appraisal of food quality.

Uml Diagram: The organic products submit the request, which is put away into an information base. The request is prepared and after completed it is refreshed into the information base. A testing is produced, and the client is sent a notice to get the request. To check for quality and status. This segment will straightforwardly utilize the product to do these capacities. The segment will be relegated to an interesting id for each product of the soil request they place, for recognizable proof. The request

segment will hold all the information given by the client. The information base part assumes a significant function in putting away, getting to, sorting out the information given by the client. This part utilizes a table with a few fields to store the information as needs be. This must be gotten to by the Admin of the shop, client segment has no admittance to it. Mistakes in this part can mess major up like losing all information, so it should consistently be kept up and upheld. Figure 1,2,3,4 and 5 represents UML diagram of various flow of activity class, sequential, flow diagram and component diagram.

Figure 1: Class diagram

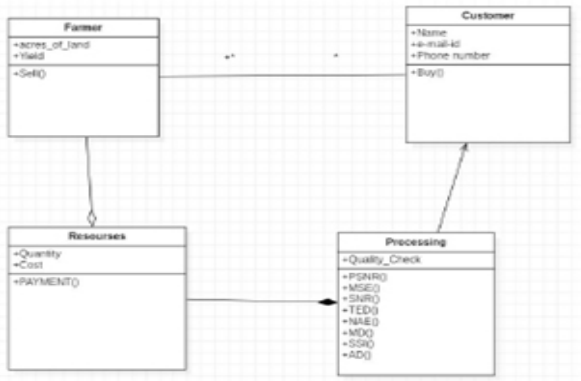


Figure 2: Sequence diagram

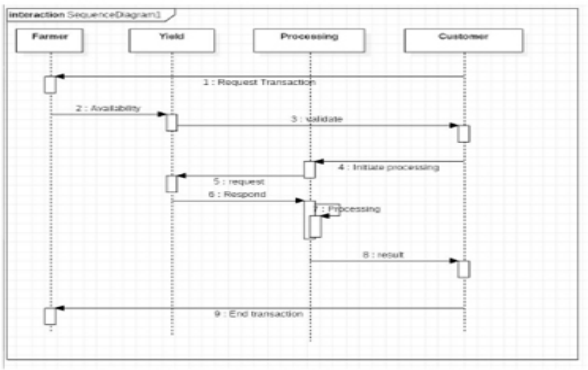
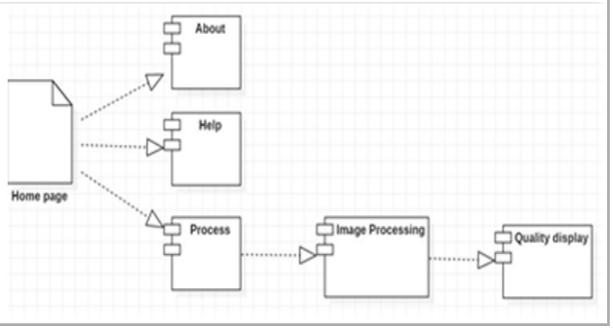


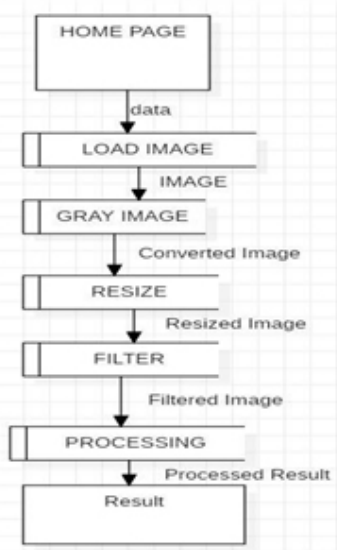
Figure 5: Component diagram



Testing Phase: Quality checking of rural items is created easy through image preparing. are often a tool that has unimaginable breadth in future. As innovation improves

Figure 3: Activity diagram

Figure 4: Data Flow diagram



work has been diminishing. Photos of the things can be tried by a progression of steps and therefore the outcomes are obtained in

Associate in Nursing exceptionally temporary timeframe. This prices less work because it decreases the work to travel to the fields to ascertain every item. This may likewise facilitate robots in future to effectively decide and get items in ranches. Imaging can be characterized because of the portrayal of an item's outer structure. That definition does not remain constant. a lot of information within an image may be thought of Future imaging frameworks are needed to be more affordable. they ought to be less complicated to utilize. the use of examining strategies and measurable investigations for picture examination are expected to separate substantial picture esteems.

Figure 6a: Test case1

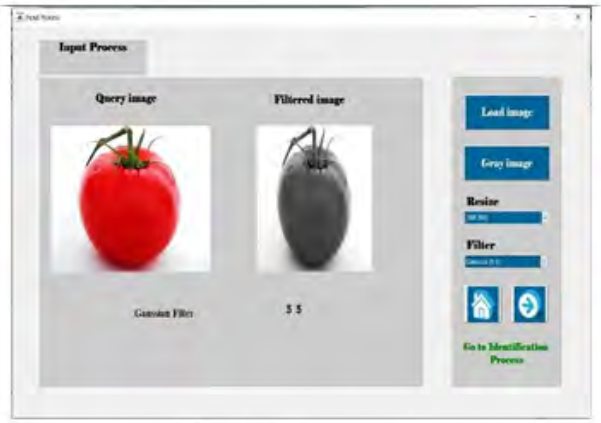
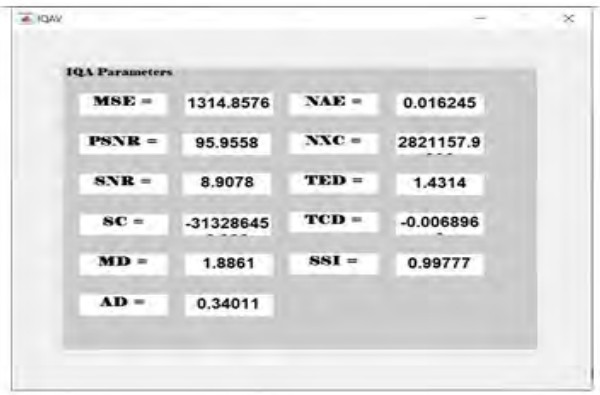


Figure 6b: Result of Test case-1using IQA parameter



Test case 2

Figure 7a: Test case2

The yield of a product getting ready can be either a product or heaps of attributes or boundaries known with the image. Most product preparing methods embody concerning the product as a two-dimensional sign and applying commonplace sign handling procedures to it. Figure 6,7 and 8 represents the three test cases are identify the Quality Assessment of Agricultural things The optimization tools was used for estimating the assimilation and diminished dissipating coefficients for an assortment of plant items and for Be that

Figure 6c: Result of Test case1

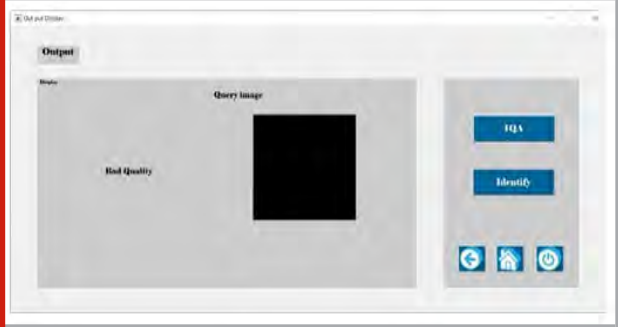


Figure 7b: Result of Test case-2using IQA parameter

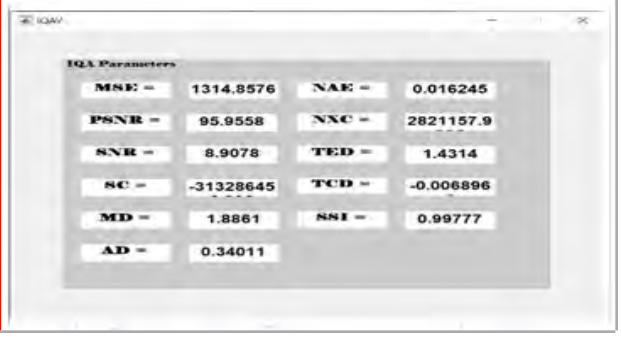
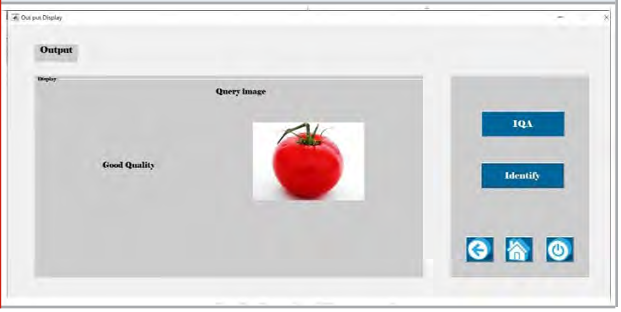


Figure 7c: Result of Test case 2



Test case 3

Figure 8a: Test case 3

because it may, the strategy needs perplexed numerical demonstrating and is inclined to mistake throughout the image getting and bend fitting.

Figure 8b: Result of Test case-3using IQA parameter

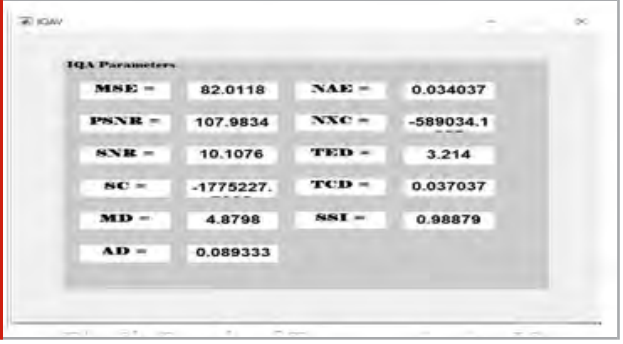
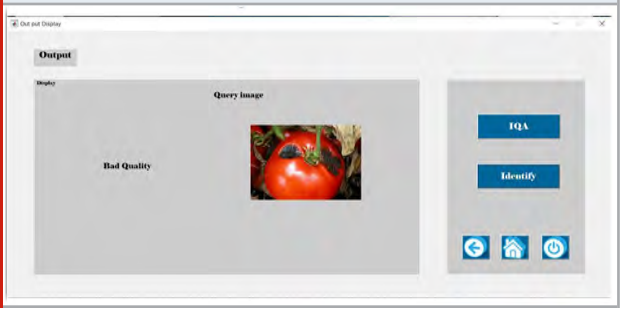


Figure 8c: Result of Test case3



Also, the strategy presently is simply affordable for center use as a result of moderate imaging and knowledge getting ready speed. additional work to boost the accuracy and speed for optical property estimation.

CONLUSION

A big take a look at for programmed image examination is that the sheer varied nature of the visual assignment that has been New innovative advancement within the territories of computerized calculation Associate in Nursing and media transmission has significance for future utilizations of picture. This innovation is required for different types of observation, measurable data assortment in the fields of customer service to identify the quality of the product therefore as to remove bad quality product to sale in market.

REFERENCES

A Fruit Quality Management System Based on Image Processing, Volume 8, Issue 6 (Nov. - Dec.2013)

Fruit Quality Management using Image Processing Mr. Sumit S. Telang, Prof. S.M.Shirsath in International Conference on Ideas, Impact, and Innovation in Mechanical Engineering, volume: 5, issue: 6 727 –733(2017)

https://in.mathworks.com/videos/introduction-to-matlab-with-image-processing-toolbox-90409.html

https://staruml

Iza Sazanita Isa, Siti Noraini Sulaiman, Muzaimi Mustapha, Sailudin Darus.Evaluating Denoising Performances of Fundamental Filters for T2-Weighted MRI Images", Procedia Computer Science, 2015.

Plant Disease Classification Using Image Segmentation and SVM Techniques K. Elangovan, S. Nalini International Journal of Computational Intelligence Research ISSN 0973-1873 Volume 13, Number 7 (2017), pp. 1821-1828

Quality measurement of fruits and vegetables, Judith A. Abbott Received 30 June 1998; accepted 11 Nov.1998

Rapid Color Grading for Fruit Quality Evaluation Using Direct Color Mapping, Published in Automation Science and Engineering, IEEE Transactions on, Volume:8, Issue: 2, 2010

Tao, Y., Heinemann, P.H., Varghese, Z., Morrow, C.T., Sommer, H.J., "III, Machine vision for colour inspection of potatoes and apples", Transactions of the ASAE, 38(5), pp.1555-1561,1995a.

XuQiabao, Zhou Xiaobo, and Zhao Jiewen, "On-Line Detection of Defects on Fruit by Machine vision Systems Based on Three Color-Cameras Systems," Computer and Computing Technologies in Agriculture II, vol. 3, pp. 2231-2238, 2009.

Biosc.Biotech.Res.Comm. Special Issue Vol 13 No 11 (2020) Pp-10-14

BBRC
Bioscience Biotechnology Research Communications

A Hybrid Approach for Autism Spectrum Disorder Classification

V. Jalaja Jayalakshmi[1] and V. Geetha[1]
[1]*Department of Computer Applications, Kumaraguru College of Technology, Coimbatore*

ABSTRACT

Autism spectrum disorder (ASD) is a neurological condition that can be devastating to the social functioning of the affected person. It is attributed to a range of symptoms that include troubles in social interaction, difficulty in expressing themselves and repetitive pattern filled behavior. People with autism have a unique behavioral pattern and the severity of the disease may vary across individuals, the causes for which are not known. The prevalence of ASD is increasing globally and early diagnosis of the disorder can lead to substantial behavioral improvements. Machine learning techniques are widely used in the health care domain for medical diagnosis. The study focuses on applying machine learning ensemble techniques to autism adult data sets to predict autism in adults. The UCI Machine Learning Repository's Autistic Spectrum Disorder Screening Data for Adult was used for the experiment purpose. The hybrid approach makes use of rough set algorithms for feature selection using Rosetta rough set tool and Adaboost with decision stump for classification using Weka data mining tool. Classification accuracy was high when the dataset was selected based on the reducts generated by Genetic algorithm. Results indicate that the proposed hybrid model improves the performance of autism data classification.

KEY WORDS: AUTISM, ENSEMBLE METHODS, MACHINE LEARNING, REDUCTS, ROUGH SET.

INTRODUCTION

Autism Spectrum Disorder (ASD) is a complex developmental disorder associated with symptoms that involve "persistent defects across multiple contexts in social communication and social interaction" and "limited, repetitive behavior, interest, or activity patterns". Autism is classified as a "spectrum" condition because the form and nature of symptoms that people encounter differ widely. Throughout all ethnic, racial, and economic groups, ASD occurs. While ASD can be a lifelong condition, the symptoms and capacity to function of an individual can be improved by therapies and services.

Diagnosis during the early stages of life significantly enhances the future of children with autism, by allowing for therapy when the brain of the child is still rapidly developing. People with autism have a unique behavioral pattern and the severity of the disease may vary across individuals, the causes for which are not known. The prevalence of ASD is increasing globally and early diagnosis of the disorder can lead to substantial behavioral improvements.

Machine learning is an application that gives systems the capability to learn and develop automatically from knowledge without being specifically programmed. Machine learning offers smart alternatives to the study of large data volumes. Machine Learning can generate precise results and analysis by designing quick and efficient algorithms and data-driven models for real-time data processing.

ARTICLE INFORMATION

*****Corresponding Author: *jalajajayalakshmi.v.mca@kct.ac.in*
Received 9th Oct 2020 Accepted after revision 23rd Nov 2020
Print ISSN: 0974-6455 Online ISSN: 2321-4007 CODEN: BBRCBA

Thomson Reuters ISI Web of Science Clarivate Analytics USA and Crossref Indexed Journal

NAAS Journal Score 2020 (4.31) SJIF: 2020 (7.728)

Online Contents Available at: http//www.bbrc.in/
Doi: http://dx.doi.org/10.21786/bbrc/13.11/3

In enhancing the efficacy of behavioral health screeners, machine learning can play a vital role (Halim Abbas et al, 2018). In several observational studies on autism data sets (Tabtah F, 2017), machine learning approaches have been commonly used. Machine learning has great capabilities to improve the analytical and intervention research in the behavioral sciences and may be especially useful in investigations concerning the highly widespread and varied syndrome of autism spectrum disorder (Bone et al, 2015). Support Vector Machines, k-nearest neighbour & Random Forest (RF) were applied on the autism data set and the results indicate that RF has got better performance in terms of accuracy measure. Weighted decision tree prediction model has been proposed for Autism Risk Analysis (Mythili et al., 2016).

Suman Raj et al (2020) strongly recommend implementing a CNN based model detection of Autism Spectrum Disorder. Accuracy of prediction of Autism Spectrum Disorder by Multilayer Perceptron classifier is better than the well-known algorithms as suggested by the authors (Jalaja et al., 2019). A study (Aboul Ella Hassanien et al., 2008) indicates that rough sets can be used for inductive learning and for constructing expert systems. Rough set theory has algorithms for knowledge reduction, concept approximation, decision rule induction, and object classification (Prerna Mahajan et al., 2012). The authors have applied a hybrid data mining model and ensemble learning classifiers to predict the credit scoring in banking domain (Koutanaei et al., 2015). Rough set-based ensemble algorithms has been used on four agricultural data sets and the results how that these algorithms perform better than the existing algorithms (Shi et al., 2018).This paper recommends a hybrid approach using rough set theory and Ada boosting algorithms for classifying the autism data set.

The organization of the paper is as follows: Section 2 describes about the algorithms used in the proposed methodology. The Experimental setup and results obtained are examined in Section 3. The concluding remarks and future work are given in Section 4.

MATERIAL AND METHODS

This work proposes a rough set-based ensemble model for effective classification of autism data set. The hybrid approach makes use of rough set algorithms for feature selection using Rosetta rough set tool and Adaboost with decision stump for classification using Weka data mining tool. The basic concepts of various algorithms used in this work are discussed below:

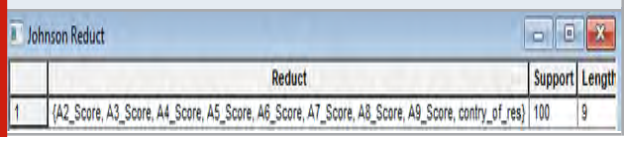
Rough Sets: Rough sets, proposed by Pawlak, is a mathematical tool for analyzing data. It has efficient procedures for finding hidden patterns of data. Rough set theory has algorithms for attribute selection and extraction, data reduction using reducts, rule generation, and pattern extraction. It stores data in tabular form, with each row representing an object. The table contains conditional attributes and a decision attribute and is known as an information system or a decision table.

If many objects in the decision table contain the same attribute values, then only one representative object is chosen, thus removing the redundant objects. These objects are known as indiscernible objects. Rough set makes use of lower and upper approximations for representation of a given set, if the available information is not adequate to decide the precise value of the set. This method of using approximation is the main idea of rough set theory. If all the objects surely belong to the set, it is known as lower approximation and if they possibly belong to the set, it is called as upper approximation.

Feature Selection: The desired features of the decision table in rough sets are selected using reducts. Reducts are the subset of conditional attributes which are adequate to categorize the decision table. The attributes that maintain the indiscernibility relation are only considered and there might be several such subset of attributes. Reducts contain the minimal attributes. The proposed work uses Johnson's and genetic algorithms for generating reducts.

Johnson's Algorithm: This algorithm is used for generating reducts using a heuristic approach. The algorithm takes the discernibility matrix as the input and counts the occurrence of each attribute. The attribute that occurs the highest number of times is added to the reduct candidate set which is initially empty. After adding the attribute to the reduct set, all the cells that contain the attribute are removed from the matrix. This process is iterated until all the non-empty cells are eliminated. This algorithm will return only a single reduct as shown in Figure 1.

FIGURE 1: REDUCT GENERATED USING JOHNSON ALGORITHM


Johnson Reduct

	Reduct	Support	Length
1	{A2_Score, A3_Score, A4_Score, A5_Score, A6_Score, A7_Score, A8_Score, A9_Score, contry_of_res}	100	9

Figure 2: Reducts Generated By Genetic Algorithm

genetic reducts

	Reduct	Support	Length
1	{A2_Score, A3_Score, A4_Score, A8_Score, A9_Score, age, contry_of_res}	100	7
2	{A1_Score, A2_Score, A4_Score, A8_Score, A9_Score, age, ethnicity, contry_of_res}	100	8
3	{A1_Score, A2_Score, A4_Score, A6_Score, A8_Score, A10_Score, age, gender, ethnicity}	100	9
4	{A1_Score, A2_Score, A4_Score, A6_Score, A7_Score, A8_Score, A10_Score, age, ethnicity}	100	9
5	{A1_Score, A2_Score, A4_Score, A6_Score, A7_Score, A8_Score, age, gender, ethnicity}	100	9
6	{A1_Score, A2_Score, A3_Score, A4_Score, A6_Score, A8_Score, A10_Score, age, ethnicity}	100	9
7	{A1_Score, A2_Score, A4_Score, A6_Score, A8_Score, A9_Score, age, gender, ethnicity}	100	9
8	{A1_Score, A2_Score, A4_Score, A6_Score, A8_Score, A9_Score, A10_Score, age, ethnicity}	100	9
9	{A2_Score, A4_Score, A6_Score, A7_Score, A8_Score, A9_Score, A10_Score, age, contry_of_res}	100	9
10	{A1_Score, A2_Score, A4_Score, A6_Score, A8_Score, A10_Score, age, ethnicity, austim, contry_of_res}	100	10
11	{A2_Score, A3_Score, A6_Score, A8_Score, A9_Score, A10_Score, age, ethnicity, austim, contry_of_res}	100	10
12	{A1_Score, A2_Score, A4_Score, A6_Score, A7_Score, A8_Score, age, ethnicity, jundice, contry_of_res}	100	10
13	{A2_Score, A4_Score, A5_Score, A6_Score, A8_Score, A9_Score, A10_Score, age, ethnicity, jundice}	100	10
14	{A1_Score, A2_Score, A4_Score, A5_Score, A6_Score, A8_Score, age, gender, ethnicity, contry_of_res}	100	10
15	{A1_Score, A2_Score, A4_Score, A6_Score, A8_Score, age, gender, ethnicity, jundice, contry_of_res}	100	10
16	{A1_Score, A2_Score, A3_Score, A4_Score, A8_Score, A10_Score, age, gender, austim, contry_of_res}	100	10
17	{A2_Score, A6_Score, A7_Score, A8_Score, A9_Score, age, gender, ethnicity, austim, contry_of_res}	100	10
18	{A1_Score, A2_Score, A4_Score, A5_Score, A8_Score, age, ethnicity, jundice, austim, contry_of_res}	100	10
19	{A2_Score, A4_Score, A6_Score, A7_Score, A8_Score, A9_Score, age, ethnicity, jundice, austim, contry_of_res}	100	11

Genetic Algorithm: Genetic algorithms are methods of optimization based on the concepts of evolution. This is a stochastic approach focused on the mechanics of natural genetics and biological evolution for function optimization. To generate better approximations, genetic

algorithms function on a population. A new population is produced at each generation by selecting individuals in the problem domain according to their level of fitness and recombining them together using operators borrowed from natural genetics. This algorithm is used to generate the best possible set of reducts for the Autism Adult data set and is shown in the Figure 2.

Adaboost: Ensemble methods is a powerful machine learning paradigm that merges predictions from several models to boost the performance of classification algorithms. Bagging and boosting are ensemble methods in machine learning algorithms. Decision trees are one of the widely used classification algorithms but tend to suffer from bias and variance Ada boost in an iterative ensemble method which combines several weak classifiers to build a single strong classifier. It initially builds a model from the training data. Then it creates a next model by correcting the errors in the first model. The process is repeated until the training set predicts correctly or the maximum model gets added. The various versions of the model use the technique of bootstrapping in which the samples are selected with replacement.

RESULTS AND DISCUSSION

Data Description: The UCI Machine Learning Repository's Autistic Spectrum Disorder Screening Data for Adult (Dua D and Graff C, 2019) was used for the experiment purpose. The dataset consists of 21 variables with 10 behavioural features, 10 individual characteristics and one class variable denoting if the person has Autism or not.

Data Pre-processing: The dataset consists of 704 instances. The data is preprocessed to enhance the quality of the data. The data cleaning step is used to remove noise and correct the inconsistencies in the data. In the dataset, the variables 'ethnicity' and 'relation' both had missing values in 95 records. 86 out of 95 are with class 'NO' and 9 out 95 are with class 'YES'. Since the variables with missing values are of categorical type, it is difficult to replace the missing value. Hence the 95 records were removed.

A record with outlier in the 'age' column was found with the value 383, which is not possible. Considering it as a typo error, the age with 383 was replaced with 38. After data cleaning, out of the 609 records, 180 records were with class value 'yes' and 429 records were with class value 'NO'. The variables 'age_desc' and 'used_app' were removed since these had the same value in all the records. After replacing the missing values and removing the unwanted attributes, rough set concepts from the Rosetta rough set tool was used for discretization and feature selection. The process of translating the values of continuous data attributes into a limited set of intervals and associating certain data values with each interval is data discretization.

Rosetta rough set tool has various algorithms for discretization and equal frequency binning algorithm was chosen to be applied on the autism data set. Equal frequency binning algorithm splits the data into 'k' number of bins and all the bins will have equal number of values. Once discretization is complete, feature selection methods are applied on the data set. These methods are used to increase the performance of classifier algorithms. In rough set theory feature selection is carried out by the generation of reducts using rough set algorithms. The entire framework for data preprocessing is shown in Figure 3.

Ensemble methods on Autism data set: Ensemble methods have been used on autism adult data set for finding decision rules to predict the occurrence of Autism Spectrum Disorder using Weka data mining tool. The attributes from the data set were selected based on the reducts generated by the rough set algorithms. The updated data set was randomly split into training and test data using varying split factors. Then Ada boosting algorithms with decision stump were used on the training data set. The rules generated were then validated with the test data set. The number of rules generated by genetic algorithm is much larger than that generated by Johnson's algorithm. The steps in the application of ensemble methods on the decision tree is shown in Figure 4.

Figure 3: Framework For Data Preprocessing

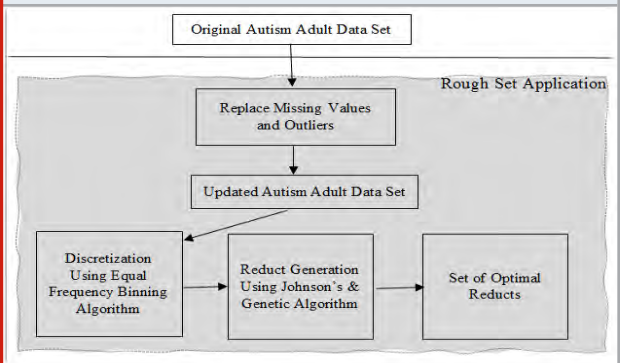
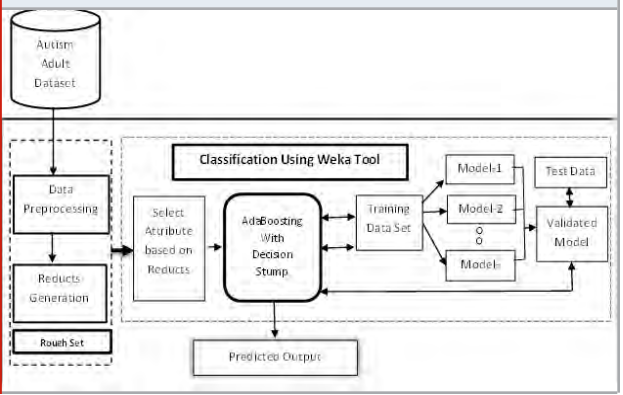


Figure 4: Framework Of The Proposed Approach



Evaluation of the Proposed Model: The performance of the model was evaluated in terms of accuracy, Kappa Statistics and F-measure. The results obtained using Ada boosting algorithms using the attributed selected from

reducts generated by Johnson's algorithm and Genetic algorithms were compared with the application of the same algorithm on the data set without using reducts and the results are discussed below.

Accuracy: Accuracy refers to the correctly classified instances for the given test data set and the results are shown in Figure 5.The results indicate when the training samples are less (10% & 20%), the performance of Adaboost algorithm without feature selection is better but when the number of training samples increase (30 % to 80%), the reduction of attributes has improved the prediction accuracy. When compared to Johnson's algorithm, the genetic algorithm performs better. This is because genetic algorithms generate a greater number of reducts than Johnson's algorithm.

Figure 5: Accuracy Percentage Of Different Classifiers

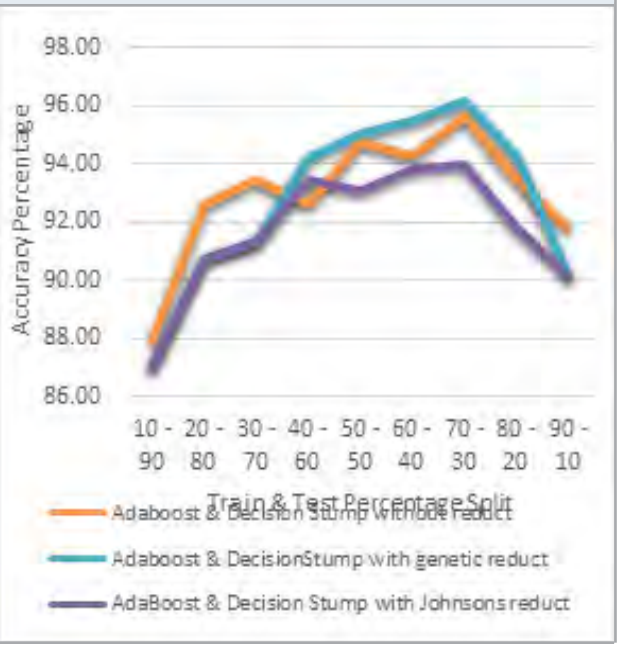
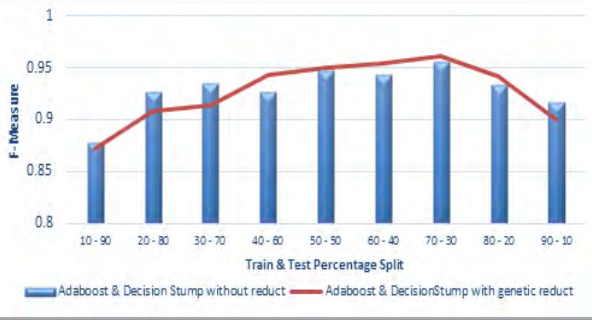


Table 1. Kappa Statistics Of Classifiers

Training-test data split percentage	Adaboost & Decision Stump without reduct	Adaboost & Decision Stump with genetic reduct
10 - 90	0.702	0.696
20 - 80	0.829	0.778
30 - 70	0.846	0.795
40 - 60	0.828	0.864
50 - 50	0.880	0.886
60 - 40	0.863	0.889
70 - 30	0.889	0.902
80 - 20	0.847	0.867
90 - 10	0.822	0.788

Figure 6: Comparison of F-Measure Values Between The Two Classifiers



Kappa Statistics: Kappa Statistic is a measure of the agreement between the predicted and the actual classifications in a dataset and the values lie between -1 to 1. A Kappa statistic value greater than 0.7 is usually considered as a good statistic correlation, but higher the value, the better the correlation. Table 1 helps to visualize the interpretation of Kappa value. It can be clearly inferred from Table.1. that using reducts has the highest Kappa statistic 0.90 and the lowest 0.696, which indicates a high to moderate degree of acceptance. The Kappa value ranges from 0.702 to 0.889 when reducts are not used. The agreement of prediction given by Kappa statistic is highest when Adaboost algorithm is applied on the reduced data set, while it is the lowest in dataset without reducts.

F-Measure: F-measure is a weighted combination of precision and recall. It finds the proportion of the true positives to the mean of predicted positives and real positives. The F-measure blends precision and recall into a single performance measure. The F-measure value varies from 0 to 100% and a higher F-measure value implies a better quality of the classification. A system with high F-measure has good precision and recall. The results based on F-measure with and without using reducts are compared and is shown in Figure 6. The quality of the best F-measure result of Adaboost with reduct (0.961%) is achieved for 70% training samples. This is higher than the algorithm without using reducts by 0.5%. This indicates that there is not much significant difference of F-measure values between both the algorithms.

CONCLUSION & FUTURE WORK

A new hybrid model has been designed for Autism detection using Rough set feature selection and ensemble learning classifiers. Two feature selection algorithms of rough set theory were applied to obtain a set of suitable subset of features, which gives a good classification performance when Adaboost learning classifier was used. Among the feature selection algorithms, genetic algorithm has given a better performance. Classification accuracy was high when the dataset was selected based on the reducts generated by Genetic algorithm. The proposed hybrid model can be used to predict autism spectrum disorder at an early stage to provide an immediate treatment. As a part of future work, other

feature selection algorithms can be applied on the data set and tested.

REFERENCES

Abbas H Garberson F Glover E and Wall DP (2018) Machine Learning Approach for Early Detection of Autism by Combining questionnaire and home video screening, Journal of the American Medical Informatics Association, 25(8) pp.1000–1007.

Bone D et al., (2015) Applying Machine Learning to Facilitate Autism Diagnostics: Pitfalls and Promises, J Autism Dev Disorf 45(5), pp.1121-1136.

Dua D and Graff C (2019) UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Hassanien AE Abdelhafez ME and Own HS (2008) Rough Sets Data Analysis in Knowledge Discovery: A Case of Kuwaiti Diabetic Children Patients, Advances in Fuzzy Systems, Volume 2008, pp.1-13.

Jalaja Jayalakshmi V Geetha V and Vivek R (2019) Classification of Autism Spectrum Disorder Data using Machine Learning Techniques, International Journal of Engineering and Advanced Technology (IJEAT), ISSN: 2249 – 8958, Volume-8 Issue-6S, pp.565-569.

Koutanaei FN Sajedi H and Khanbabaei M (2015) A Hybrid Data Mining Model of Feature Selection Algorithms and Ensemble Learning Classifiers for Credit Scoring, Journal of Retailing and Consumer Services, vol. 27, pp.11-23.

Mahajan P, Kandwal R and Vijay R (2012) Rough Set Approach in Machine Learning: A Review, International Journal of Computer Applications 56(10), pp.1-13.

Mythili MS and Mohamed Shanavas AR (2016) An Improved Autism Predictive Mechanism among Children Using Fuzzy Cognitive Map and Feature Extraction Methods (Feast), ARPN Journal Of Engineering And Applied Sciences, Vol. 11, No. 3, pp.1451-1456.

Raj S and Masood S (2020) Analysis and Detection of Autism Spectrum Disorder Using Machine Learning Techniques", Procedia Computer Science ,167, pp.994–1004.

Shi L Duan Q Zhang J Xi L and Ma X (2018) Rough Set based Ensemble Learning Algorithm for Agricultural Data Classification, Filomat 32(5), pp.1917–1930.

Tabtah F (2017) Autism Spectrum Disorder Screening: Machine Learning Adaptation and DSM-5 Fulfillment", Proceedings of the 1st International Conference on Medical and Health Informatics 2017, Taichung City, Taiwan, ACM., pp.1-6.

Biosc.Biotech.Res.Comm. Special Issue Vol 13 No 11 (2020) Pp-15-19

BBRC
Bioscience Biotechnology Research Communications

Alzheimer Disease Forecasting using Machine Learning Algorithm

Malavika G[1], Rajathi N[2], Vanitha V[3] and Parameswari P[4]

[1]*PG Scholar, Department of Information Technology, Kumaraguru College of Technology, Coimbatore, India.*
[2,3]*Professor, Department of Information Technology, Kumaraguru College of Technology, Coimbatore, India.*
[4]*Assistant Professor (SRG) Department of MCA, Kumaraguru College of Technology, Coimbatore, India.*

ABSTRACT

Alzheimer disease is a neurodegenerative disease that makes a gradual disorder of human brain cells and it leads to degenerate the cells away and die. In India more than one million cases per year are affected by this disease. The most common in people over the age group of above 65. There is no treatment for this disease to cure, but now a day's medications are available to temporarily decline the process of disease. The primitive detection of this disease may help the doctors, physician, and other family members to treat them in a better way. The objective of the proposed system is to offer a fast, early and cost-efficient method to detect disease in premature period. Machine learning is the blooming field in the healthcare industry, so by using the machine learning techniques the disease will get forecast in the earlier stage. The techniques are K-Nearest Neighbor, Adaboost Classifier, Support Vector Machine, Logistic Regression, Decision Tree Classifier and Random Forest classifier. Among these algorithms, the best prediction accuracy is produced by the Random Forest algorithm.

KEY WORDS: ACCURACY, ALZHEIMER'S DETECTION, MACHINE LEARNING, PRIMITIVE DETECTION.

INTRODUCTION

Alzheimer's disease is a cause of dementia. Dementia leads to memory loss, thinking ability due to some of the brain disease. Alzheimer is one of the brain diseases that causes Dementia. This disease causes the mini strokes in the brain and that occurs the gradual cell destruction and the nerve disorder in the brain. A person who is affected by the disease may not be aware of the strokes due to the minor attacks and that occurs without any perception. It occurs at individual losses.

ARTICLE INFORMATION

*****Corresponding Author: rajathi.in.it@kct.ac.in
Received 8th Oct 2020 Accepted after revision 26th Nov 2020
Print ISSN: 0974-6455 Online ISSN: 2321-4007 CODEN: BBRCBA

Thomson Reuters ISI Web of Science Clarivate Analytics USA and Crossref Indexed Journal

NAAS Journal Score 2020 (4.31) SJIF: 2020 (7.728)

Online Contents Available at: http//www.bbrc.in/
Doi: http://dx.doi.org/10.21786/bbrc/13.11/4

This disease mainly affects the age of 65, it is not possible to calculate that age nowadays, it can occur as early as 50 unfortunately, but the early 50 cases are rare then the 65 above. The people who are affected early are usually aware of the changes in them. Their new deviations and memory loss affect them deeply, and they always forget the things and they are not able to handle their things as when they are in normal condition. They feel some difficulty to talk and to use the words, while talking with family members, relatives, friends etc.... This leads them to talk less and this advanced stage leads to forgetting the close family members. When they release that they are not functioning as well as they did formerly, they become depressed.

Overall analysis the maximum Americans suffer from AD's. There are 4.5 million American people suffered by this disease. Research says that these will get expected to increase in the year of 2050 as 14 million. The diagnosis of Alzheimer's disease takes a long process that has an awful effect on the patients with the disease and

their families. Analysis of AD's is not simple and easy. It cannot be done without any proper study of brain sample tissues.

There is no proper treatment for this disease to cure, it may reduce the decline but not cure the disease. If the early detection of the disease is done that will be helpful for the physician, family members all other close to them etc. So, that the Machine learning techniques are used to diagnose the disease earlier. There are five techniques used here to find the best accuracy. The techniques are K-Nearest Neighbor, Adaboost Classifier, Support Vector Machine, Logistic Regression, Decision Tree Classifier, Random Forest classifier. Among this the best and high accuracy detector can be determined by using python code implementation.

MATERIAL AND METHODS

Various studies with respect to diagnosis of Alzheimer's disease are discussed. Joshi, S, et.al., uses a various Machine Learning algorithm to categorize the AD's (Alzheimer' disease) and PD's (Parkinson's disease) with high accuracy classifier, by using the major risk factor. By using Fluorodeoxyglucose, Positron Emission Tomography and Pittsburg Compound B imaging techniques Illan, I.A, et.al., compare the forecast accuracy on early AD's (Alzheimer's Disease). Image-based classification method is used by Dong Hye Ye et.al., to classify the brain MRI scans with MCI. A semi-supervised classifier patterns are used to achieve a high sensitivity.

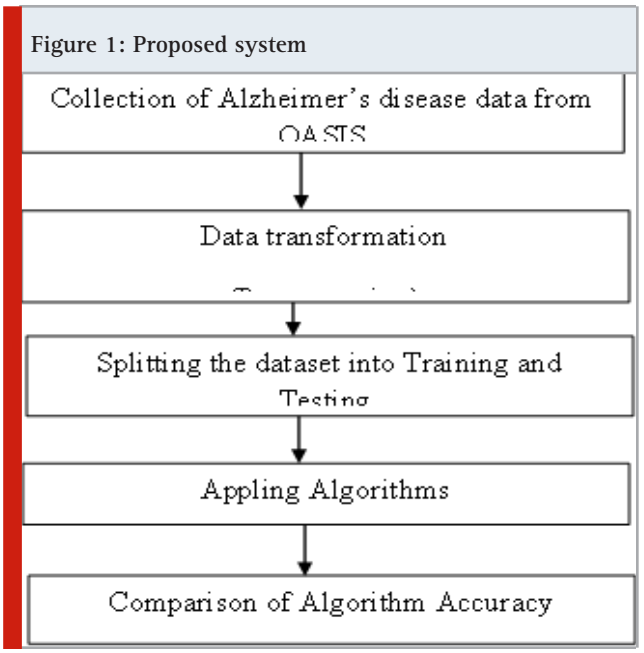
Data mining techniques are used in Aunsia Khan et.al., proposed system. They present an assessment, study and estimation of the current work in the initial detection of Alzheimer using Machine Learning algorithm. Ammarah Farooq et.al., proposed a four-way classifier to mainly classify the AD, MCI, LMCI and healthy persons. Core focus of this work is to classify the stages of disease. By using the deep learning technique. Machine Learning algorithms stand to detect the AD in Arpita Raut et.al.,. This planned method abstracts the surface and figure topographies from the MRI scans. Neural Network is used for detection of various stages of AD's. Karl Backstrom et.al., uses a deep convolutional neural network to offer an effective and simple 3D convolutional network architecture to achieve a high performance of AD detection.

The Deep Convolutional Neural Network is used for detecting Dementia and AD from MRI scan in H. M. Tarek Ullah et.al.,. This paper discussed the fast, costs less and more reliability. Priyanka Lodha et.al., paper mainly focused on using the neuroimaging techniques like CT, MRI, PET, EEG data, to detect Alzheimer in its primitive stage using ML. The assessment of ML Technique is done in Mohamed Mahyoub et.al., proposed system. They rank AD risk factors on clinical assessment data. Mohamed Mahyoub et.al., investigates five different classifiers in the risk factor of AD's data for better accuracy. Gokce UYSAL et.al., evaluate the early forecast of dementia in AD by using machine learning algorithms. Here, they consider the hippocampus brain region of dementia. This approach can be useful for separating the patients with AD and CN. The age of the brains of individuals can be very useful in many applications. Masoumeh Siar et.al., has greatly paid to forecasting and avoiding early deaths in the medical field. This paper has been done by using (DL).

J. Neelaveni et.al., used machine learning algorithms to guess Alzheimer disease using psychological parameters like age, number of visits, MMSE and education. By using the parameters as input the algorithms are applied. The SVM and Decision Tree algorithm are used and the comparison is done by the accuracy. The best accuracy detector is SVM. Aakash Shah et.al., comprises the complete study and precision of various ML techniques. Voting Classifier Algorithm is used for early discovery of Alzheimer Disease, and to removes the possibility of inaccuracies in the result. Rajathi et. al successfully applied machine learning methods for disease prediction.

Proposed Methodology: Machine learning plays a significant title role in the health care business. There is a large amount of database provided by the healthcare domain to develop an advanced and scientific method to diagnose the disease in an early stage. So, here some of the machine learning algorithms are used to forecast the disease and to find the best accuracy provider among these algorithms. The algorithms are Logistic Regression with null values, Logistic Regression without null values, Support vector machine, Decision Tree, Adaboost. The python code is used for the implementation. The proposed system is pictorially represented in Figure I.

Figure 1: Proposed system



Dataset: The Alzheimer's disease dataset is collected from OASIS, which is offered on their website. It can be applied and used for the purpose of training and

executing the algorithms to identify the disease impact. Here, the longitudinal MRI data are used. The dataset consists of a longitudinal MRI data of 150 subjects aged 60 to 96.

Data pre-processing: Data pre-processing is done to remove the rows with missing values, splitting Training and Testing data and to cross validate the data.

Figure 2: Groupwise classification.

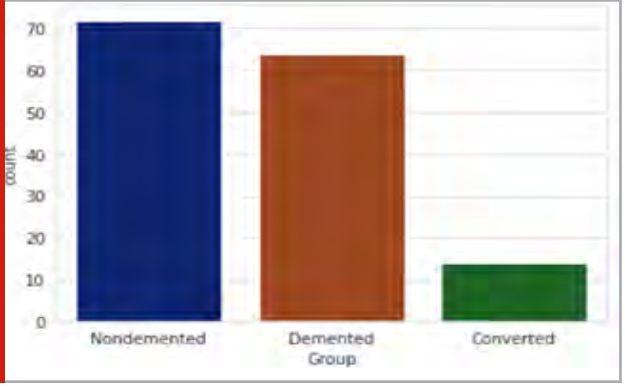


Figure 3: Converted cases into demented

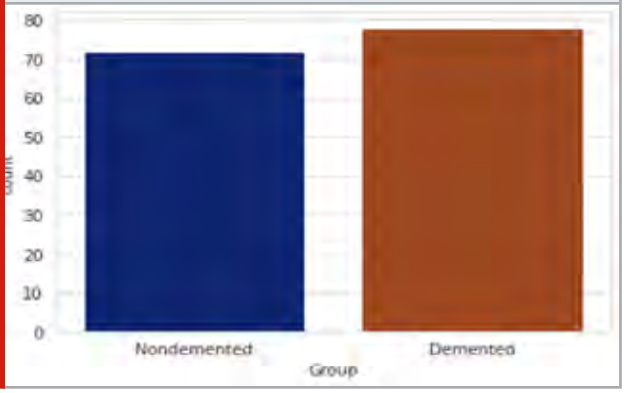
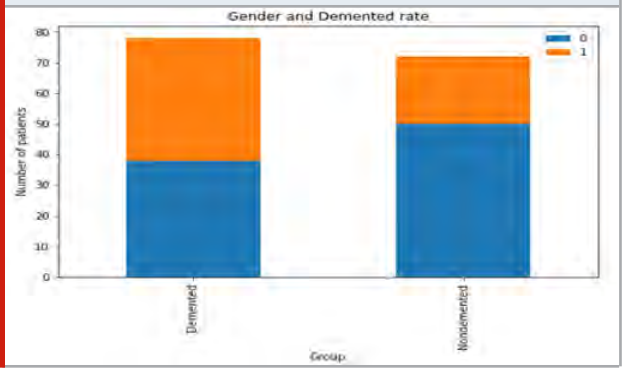


Figure 4: Presence of disease based on Gender



RESULTS AND DISCUSSION

The overall objective of this paper is to predict more accurately the early detection of Alzheimer disease. By using technologies named K-Nearest Neighbor, Adaboost Classifier, Support Vector Machine, Logistic Regression, Decision Tree Classifier, Random Forest classifier. From the result it's been seen that the Random forest and Adaboost gives more accuracy as compared as other techniques.

Figure 5: Age vs count.

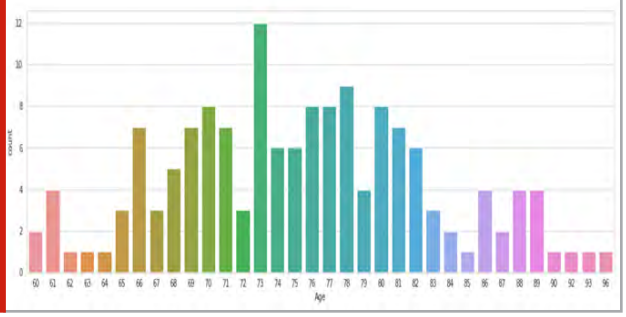
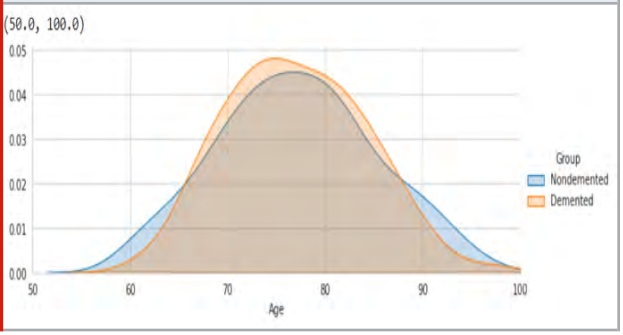


Figure 6: Age vs Group



The fig -2 shows that the number of Non demented, Demented and Converted cases. It clearly shows that Nondemented are higher when compared to other, the demented group be less when compared. The converted shows as demented, they are Non demented to Demented case. Fig 3 shows the converted cases into demented. The above Fig- 4 clearly says that the difference between the Demented and Non-Demented patients with respective Gender (0= Female(blue), 1= Male(orange)).

This clearly shows that greater number of are in Demented condition as compared to female. In Non-Demented Condition Females are higher than Male. The fig-5 shows that the maximum age of diseased cases. The age between 68 to 83 be the most affected case in Alzheimer's disease. Fig.6 shows that the gradual increase in 68 age and it's in the peak of mid-70 to 80 and it gradually decreased in the age of after 80.

The performance of the classification models on the test data was represented in Table – 1. The various performance parameters Precision, Recall, F1 scores for both male and female of various models are presented in Table 2.

Table 1. Performance of Algorithms

Methodology	Precision		Recall		F1 Score	
	Female(0)	Male(1)	Female(0)	Male(1)	Female(0)	Male(1)
Logistic Regression Decision Tree	0.69	0.79	0.79	0.70	0.74	0.74
Classifier	0.75	0.84	0.83	0.77	0.79	0.80
K-Nearest Neighbor	0.61	0.77	0.81	0.55	0.69	0.64
Support Vector Machine	0.71	0.87	0.88	0.68	0.79	0.77
AdaBoost	0.79	0.82	0.79	0.82	0.79	0.82
Random Forest	0.80	0.88	0.87	0.82	0.83	0.84

Table 2. Classification Performance

Methodology	Classification Accuracy
Logistic Regression	74.1%
Decision Tree Classifier	79.4%
K-Nearest Neighbor	66.9%
Support Vector Machine	77.6%
AdaBoost Classifier	80.3%
Random Forest Classifier	86.8%

From the results obtained, the Random Forest classifier gives the high accuracy than the other models

CONCLUSION

In this paper, various machine learning algorithm were used to predict the Alzheimer disease at early stage. The results obtained shows that the Random forest classifier gives the best performance when compared to other methods. The future work is to apply hybrid approaches and their performances to be studied.

REFERENCES

Aakash Shah, Dhruvi Lalakiya, D., Shekha Desai., Shreya, and Vibha Patel., 2020, June. Early Detection of Alzheimer's Disease Using Various Machine Learning Techniques: A Comparative Study. In 2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184) (pp. 522-526). IEEE.

Aunsia Khan. and Muhammad Usman., 2015, November. Early diagnosis of Alzheimer's disease using machine learning techniques: A review paper. In 2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K) (Vol. 1, pp. 380-387). IEEE.

Ammarah Farooq., Syed Muhammad Anwar., Muhammad Awais and Saad Rehman., 2017, October. A deep CNN based multi-class classification of Alzheimer's disease using MRI. In 2017 IEEE International Conference on Imaging systems and techniques (IST) (pp. 1-6). IEEE.

Arpita Raut and Vipul Dalal., 2017, July. A machine learning based approach for detection of Alzheimer's disease using analysis of hippocampus region from MRI scan. In 2017 International Conference on Computing Methodologies and Communication (ICCMC) (pp. 236-242). IEEE.

Dong Hye Ye., Pohl, K.M. and Davatzikos, C., 2011, May. Semi-supervised pattern classification: application to structural MRI of Alzheimer's disease. In 2011 International Workshop on Pattern Recognition in NeuroImaging (pp. 1-4). IEEE.

Gokce UYSAL and Mahmut OZTURK., 2019, October. Using Machine Learning Methods for Detecting Alzheimer's Disease through Hippocampal Volume Analysis. In 2019 Medical Technologies Congress (TIPTEKNO) (pp. 1-4). IEEE.

H. M. Tarek Ullah., Dr. Dip Nandi., and Zishan Ahmed Onik., 2018, April. Alzheimer's Disease and Dementia Detection from 3D Brain MRI Data Using Deep Convolutional Neural Networks. In 2018 3rd International Conference for Convergence in Technology (I2CT) (pp. 1-3). IEEE.

Illan, I.A., Gorriz, J.M., Ramirez, J., Chaves, R., Segovia, F., López, M., Salas-Gonzalez, D., Padilla, P. and Puntonet, C.G., 2010. Machine learning for very early Alzheimer's disease diagnosis; a 18 F-FDG and pib PET comparison. In IEEE Nuclear Science Symposuim & Medical Imaging Conference (pp. 2334-2337). IEEE.

Joshi, S., Shenoy, D., GG, V.S., Rrashmi, P.L., Venugopal, K.R. and Patnaik, L.M., 2010, February. Classification of Alzheimer's disease and Parkinson's disease by using machine learning and neural network methods. In 2010 Second International Conference on Machine Learning and Computing (pp. 218-222). IEEE.

Karl Bäckström., Mahmood Nazari., Irene Yu-Hua Gu and Asgeir Store Jakola., 2018, April. An efficient 3D deep convolutional network for Alzheimer's disease

diagnosis using MR images. In 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018) (pp. 149-153). IEEE.

Mahyoub Mohamed., Martin Randles., Thar Baker and Po Yang., 2018, June. Effective Use of Data Science Toward Early Prediction of Alzheimer's Disease. In 2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS) (pp. 1455-1461). IEEE.

Mohamed Mahyoub., Dr. Martin Randles., Dr. Thar Baker and Dr. Po Yang.,., 2018, September. Comparison analysis of machine learning algorithms to rank alzheimer's disease risk factors by importance. In 2018 11th International Conference on Developments in eSystems Engineering (DeSE) (pp. 1-11). IEEE.

Masoumeh Siar and Mohammad Teshnehlab, M and K.N. Toosi., 2019, October. Age Detection from Brain MRI Images Using the Deep Learning. In 2019 9th International Conference on Computer and Knowledge Engineering (ICCKE) (pp. 369-374). IEEE.

Neelaveni, J. and Geetha Devasana, M.S., 2020, March. Alzheimer Disease Prediction using Machine Learning Algorithms. In 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS) (pp. 101-104). IEEE.

Priyanka Lodha., Ajay Talele and Kishori Degaonkarm, 2018, July. Diagnosis of Alzheimer's Disease using Machine Learning. 2018 Fourth International conference on Computer Communication Control and Automation (ICCUBEA).

Rajathi, N., Kanagaraj, S., Brahmanambika, R. and Manjubarkavi, K., 2018. Early detection of dengue using machine learning algorithms. International Journal of Pure and Applied Mathematics, 118(18), pp.3881-3887.

Biosc.Biotech.Res.Comm. Special Issue Vol 13 No 11 (2020) Pp-20-23

BBRC
Bioscience Biotechnology Research Communications

Face Generation using Deep Convolutional Generative Adversarial Neural Network

Devaki P.[1]*, Prasanna Kumar, C.B[2], Kaviraj S[3] and Ramprasath A[4]
[1]*Professor, [2,3,4]UG Scholars*
[1,2,3,4]*Department of CSE, Kumaraguru College of Technology, Coimbatore, India*

ABSTRACT

Due to the huge availability of data, it is difficult to classify/process images at a higher speed and accuracy. The first technique was in the field of computer vision and it used image data for face recognition and detection of an object from the image but later Convolutional Neural Networks (CNNs) took place. CNNs are used for feature detections by looking at the image and try to check if certain features are present in the image and then it classifies the image accordingly. Acquiring and processing the dataset for the Machine learning technique is one of the time-consuming processes, so Generative Adversarial Neural Network (GAN) are introduced. GAN typically work with image dataset but they are difficult to train. This paper explores the potential of GAN to generate realistic images. Deep Convolutional Generative Adversarial Networks (DCGAN) is used to generate new images that are not in the dataset. DCGAN has a great success in generating the new images. MNIST (Modified National Institute of Standards and Technology dataset) contains images of handwritten digits dataset and CelebA dataset contains images of celebrities are used, performing the adversarial learning on it and try to generate new images as same as the MNIST and CelebA datasets.

KEY WORDS: CELEBA, CONVOLUTIONAL NEURAL NETWORKS, DEEP CONVOLUTIONAL ADVERSARIAL NEURAL NETWORKS, MNIST.

INTRODUCTION

Machine directed algorithms in the form of Machine Learning were built known as Self-Organizing Networks (SON) (Hughes et al,.2019). The SON was used to train highly co-relevant data. With the scarcity of real labeled data alarming a concern, a need for a new method arose where data was not pure, not easy to segregate, and unsupervised. Hence, a data-driven (model-free) approach built using two interconnected Artificial Neural Networks called Generative Adversarial Networks which filled the setbacks in SON. This method was able to generate synthetic data that was created by augmenting real data with realistic synthetic data. GANs were able to develop new datasets on historical data without specifying a model or fitting probability distribution.

With GANs being a promising technology in the field of machine learning approach, it played a significant role in non-labeled data (i.e.) data where the details were not sufficient. Thus, the usage of GANs became more popular in the field of semi-supervised and unsupervised learning. (Li et al,. 2018) shows that the relation of specific selections of high-level textual representations in linear models was substantial, and they failed to capture complex distributions. An observation of limitation in scalability and inference accuracy proved to be a critical factor in the failure of Gaussian and Non-parametric density models. Therefore, for the joint optimization of the model and variational parameters, stochastic inference algorithms were used. GANs provided an improvement in

ARTICLE INFORMATION

**Corresponding Author: devaki.p.cse@kct.ac.in*
Received 5th Dec 2020 Accepted after revision 27th Nov 2020
Print ISSN: 0974-6455 Online ISSN: 2321-4007 CODEN: BBRCBA

Thomson Reuters ISI Web of Science Clarivate Analytics USA and Crossref Indexed Journal

NAAS Journal Score 2020 (4.31) SJIF: 2020 (7.728)

Online Contents Available at: http//www.bbrc.in/
Doi: http://dx.doi.org/10.21786/bbrc/13.11/5

the learning of the classifiers by incorporating adversarial objectives that made it more robust and efficient.

(Ledig et al,. 2017) Application of GAN's plays a role in converting the image from Low Resolution (LR) image to High Resolution (HR) image. High MSE results in Peak Signal to Noise Ratio (PSNR), but they often lack high-frequency details and perceptual unsatisfying. Structural Similarity Index (SSIM) and PSNR are the metrics used to determine the quality of the image, and MSE is the error calculation commonly used in encoding and decoding of the image. The MSE based solution appears overly smooth image due to the pixel-wise average of possible solutions in the pixel space. At the same time, GAN drives the reconstruction towards the original image manifold producing perceptually more convincing solutions. Since MSE struggles to handle the uncertainty in recovering the high-frequency details such as texture, a loss function based on the Euclidean distance between feature maps are used to provide perceptually more satisfying result.

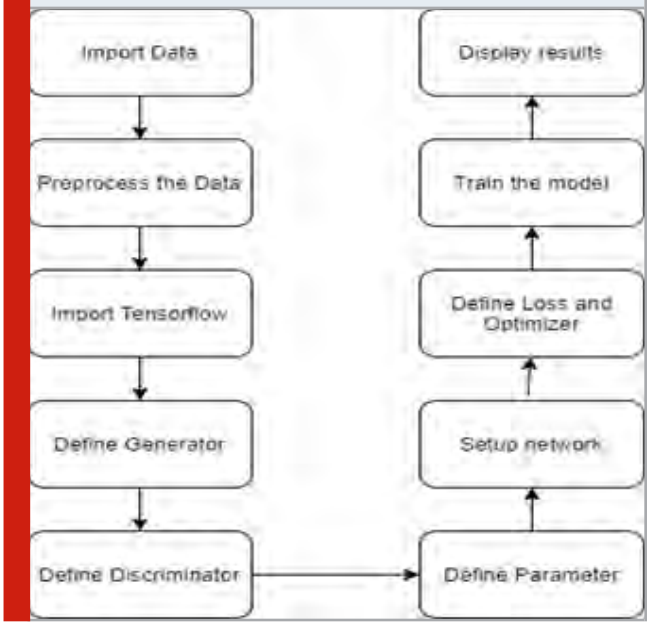
In order to obtain more efficient and clean images, batch-normalization of layers and ParametricReLU, which uses LeakyReLU as activation function, are used to obtain a sigmoid function in the two deep neural networks used respectively. This method is not optimized for video and can be used only for images. A special GAN called the Deep Convolutional Generative Adversarial Network (DCGAN) finds its usage in unsupervised data. The DCGAN is a combination of two deep artificial neural networks called the Generator and Discriminator. (Liu X et al,.2018) Uses "Age-DCGAN," where input is the image of a person's face, which is passed to the encoder and learn its personality and age features. The Discriminator network uses conditions to generate the corresponding aging face and arrive at information that shall act as base values in determining synthetic images. Thereby, Discriminator imposes on the Generator to discriminate input images and generated images until they are real. The method makes use of two sub-encoders to separate personality and age-features in simulating and predicting the aging of the human face. Results show that using two vectors in the Generator shall generate a more accurate and realistic image to match the input samples.

DCGAN find their scope extending to a greater area when it comes to images. (Arora et al,.2019) Inpainting is filling the neighboring pixels to remove an object from the image or removing a lousy effect like the red-eye effect or removing the watermark. However, in outpainting, there is a lot to explore, and the implementation of new techniques are needed. The sole aim of outpainting is to produce millions of new pixels needed to generate a similar one to the original image. DCGAN helps to outpaint the image recursively up to a greater extent, by obtaining larger extended realistic images. By giving contextual and perceptual information to the algorithm, the generator can generate the images by repeating Recursive painting (up to a maximum of 3 times) to output more clear and high-quality images that have a lesser pixel loss.

(Nataraj et al,.2019) Detecting Fake images using Co-occurrence matrices is GAN based technique to identify fake images such as Deep fakes, Image-to-Image transitions. It is inspired by classical Steganalysis, to detect manipulation of images. In this, Co-occurrence matrices are extracted on 3-color channels of a pixel domain and trained using CNN. The nature of ML used is similar to a Game Theory optimization problem that imbibes min-max to compete between generator and discriminator. Initially, using image residuals, these matrices were computed, passing it through several filters to obtain differences over which a classifier was built to check the authenticity of the image. However, authors provide a method of building convolutional layers over a neural network by using pixels of RGB from an image, thereby allowing the network to learn features from the matrix itself. The optimizer used is of stochastic gradient descent.

With all said, for further improvement (Salimans et.al,.2016) says that training GANs requires finding NASH equilibrium of a non-convex game with continuous, high dimensional parameters. Even though gradient descents provide the best outcomes to the present situations, they fail to converge sometimes. In order to improvise techniques such as Feature matching, Minibatch normalization, Historical averaging, Virtual batch normalization are some of the techniques the author suggests to improve GANs. Though GANs have shown promising development, it still lacks proper metrics and unstable training.

Figure 1: System Flow Diagram



MATERIAL AND METHODS

Vanilla Gans: These are the simplest of GANs. The generator and discriminator are only multi-layer perceptron. The Vanilla GANs function using stochastic

gradient descent algorithms. From (Gauthier Jon 2014) , by considering four arbitrary noise samples, images are mapped in rows, and the number of epochs it takes to generate a proper image is represented in columns and observed a high of 131 epochs, which is a considerably very high cost. They observed Mosaic patterns and discolorization in raw images and struggle in learning data when there is an overlap. Therefore, Vanilla GANs failed to overfit the training data.

Conditional GAN: These are the GANs with conditional parameters. From (Gauthier Jon 2014), observed that by considering a conditional parameter 'y' as face attributes in the generator, they train the network using the image data as 'x'. Only proper visual attributes are considered as attributes and the remaining attributes as noise. Training the network, they found considerable changes in similar faces, but it still failed to overfit the data. Though the system observed strong positive correlations with the generator, the cost of generator is improving overtime, but the discriminator keeps on failing. These methods led us to use a more efficient deep learning technique called DCGANs proposed in this paper.

Proposed Solution: Figure 1 is the proposed System Flow. The proposed solution seeks to generate artificial images using Deep Convolutional Generative Adversarial Networks (DCGAN). Having performed well with unlabeled (unsupervised) data in the past, DCGANs make a perfect option to this scenario. The model was run against two datasets:

1. MNIST dataset containing images of handwritten digits of 60,000 images
2. CelebA dataset containing 200,000 images of celebrity faces.

The designed solution is in a way that it adapts both the datasets with only the number of channels varying as 1(L) and 3(RGB), respectively. The system is constructed using Tensorflow v1.0.

Module 1(Generator): In the generator network Figure 2, along with every input, a random input (noise generated from Gaussian distribution) is added to scramble the original image, thereby generating a new image. This is performed for all the images provided as input. The generator also performs up sampling, where it combines a broader set of smaller images together to form a single large image. In this system, 2 hidden layers are present. Xavier initializer is used to initialize weights to make sure that neuron activation functions do not happen in zero or dead regions. By doing batch normalization in every layer for standardization, it also decreases computation cost by reducing the number of epochs. 'Tanh' is used as the activation function at the logits layer (Output layer of the network) as they work well along with Xavier initializers.

$$\text{Tanh}(x) = 2/1+e\text{-}2x\ \text{-}1.$$

Tanh activation function is preferred over sigmoid functions because the gradients get stronger and steeper over time.

Module 2: (Discriminator): The Discriminator network Figure 3 is the reverse of the generator structure. The Discriminator performs down sampling (i.e.) breaks down the large image received due to up sampling in the generator into smaller fragments. Similarly, 2 hidden layers are there in the Discriminator and 'Sigmoid function' as the activation function in the output layer to determine whether the generated image is real or fake.

Figure 2: The Generator Network

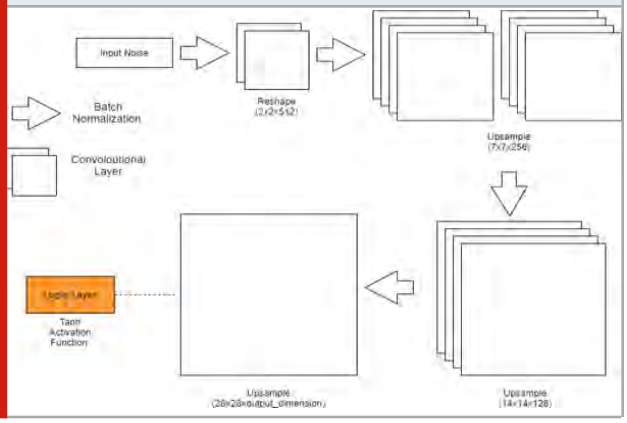


Figure 3: The Discriminator Network

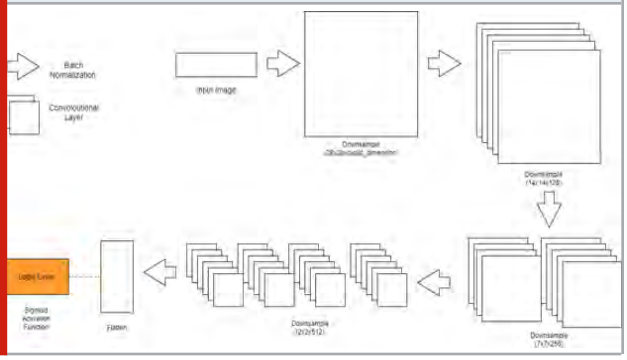


Figure 4: Sigmoid Cross-entropy loss

$$H(y,\hat{y}) = -\frac{1}{M}\sum_{j=1}^{M}[y_j \log \hat{y}_j + (1-y_j)\log(1-\hat{y}_j)]$$

$$\hat{y}_j = \sigma(z_j) = \frac{1}{1+e^{-z_j}} \text{ for score } z_j$$

Module 3: (Loss And Optimization): Generator and Discriminator loss computed are using Sigmoid cross entropy Figure 4 from the computed logits.

In Figure 4, M is the number of classes. z subscript j is the raw output of the convolutional network. yhat subscript

j is the predicted probabilities calculated using sigmoid function applied to the z. y subscript j is the ground truth of the input images.

The implementation of cross entropy is done with reference to https://github.com/carpedm20/ DCGAN -tensorflow/blob/master/ops.py#L35 Adam Optimizer, is used to provide better individual learning rates for each parameter within the network. The tolerance level of loss of generator is ensured that is lesser than the discriminator loss or close to 0.

RESULTS

By using DCGAN, after every 10 batches, the generator and discriminator loss values is observed. The generator output is also observed for every 100 batches. With batch size as 64, the MNIST dataset was trained with two learning rates 0.001 and 0.0005. Figure 5 and Figure 6 shows the outputs for the learning rates 0.001 and 0.0005 respectively. The network is trained for 2 epochs with each learning rates.

Figure 5: Output for MNIST dataset after 2 epoch with learning rate of 0.001

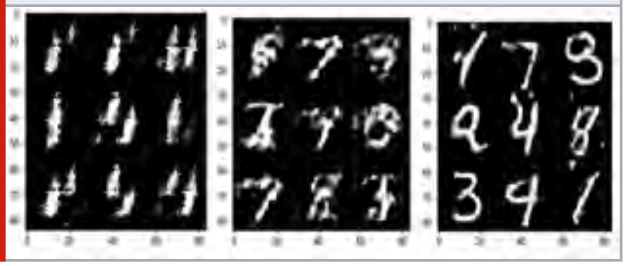
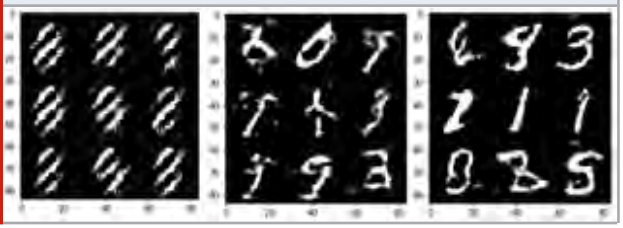


Figure 6: Output for MNIST dataset after 2 epoch with learning rate of 0.0005

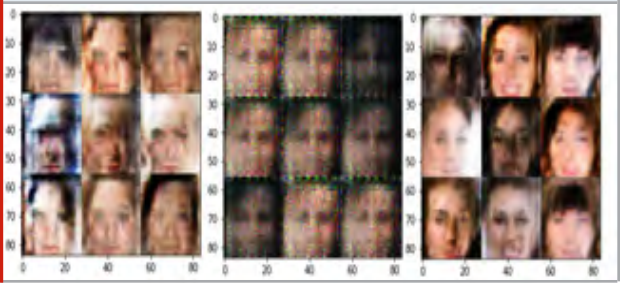


Also, the results for the CelebA dataset, which contains the images of faces is observed by setting the batch size = 64 and learning rates at 0.0005. Figure 7 is the output obtained from training the network.

CONCLUSION

By training for only two epochs, we were only able to attain the current output efficiency. By increasing the number of epochs and optimizing the neural layers and learning rate, better results could be obtained. By improving the designed network, this model shall prove to be efficient in unsupervised learning problems and can achieve state-of-the-art results with similar datasets as well.

Figure 7: Output for CelebA dataset after 2 epoch with learning rate of 0.0005



REFERENCES

Arora H Jain S Anand S and Rajpoot D S (2019) Augmentation of Images through DCGANs 12th International Conference on Contemporary Computing (IC3) Noida India pp. 1-6.

Gauthier Jon (2014) Conditional generative adversarial nets for convolutional face generation Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition, Winter semester 2014.5 (2014): 2

Hughes B Bothe S Farooq H and Imran A (2019) Generative Adversarial Learning for Machine Learning empowered Self Organizing 5G Networks International Conference on Computing, Networking and Communications (ICNC) Honolulu HI USA pp. 282-286.

Ledig C et al. (2017) Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Honolulu HI pp. 105-114.

Li T Liu X and Su S (2018) Semi-supervised Text Regression with Conditional Generative Adversarial Networks IEEE International Conference on Big Data (Big Data), Seattle WA USA pp. 5375-5377.

Liu X Xie C Kuang H and Ma X (2018) Face Aging Simulation with Deep Convolutional Generative Adversarial Networks 10th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA) Changsha pp. 220-224.

Nataraj L et al.(2019) Detecting GAN generated Fake Images using Co-occurrence Matrices

Salimans Tim Ian J et.al., (2016) Improved Techniques for Training GANs NIPS.

Biosc.Biotech.Res.Comm. Special Issue Vol 13 No 11 (2020) Pp-24-27

BBRC
Bioscience Biotechnology Research Communications

Heart Disease Prediction Using Machine Learning Algorithms

Malavika G[1], Rajathi N[2], Vanitha V[3] and Parameswari P[4]

[1]*PG Scholar, Department of Information Technology, Kumaraguru College of Technology, Coimbatore, India.*
[2,3]*Professor, Department of Information Technology, Kumaraguru College of Technology, Coimbatore, India.*
[4]*Assistant Professor (SRG) Department of MCA, Kumaraguru College of Technology, Coimbatore, India.*

ABSTRACT

The rapidly growing field of data analysis plays a significant role in healthcare. The healthcare industry has become big business. The healthcare sector produces enormous amounts of data every day. This data helps to extract the hidden information, which is useful to predict disease at the earlier. In medical field, predicting heart disease is treated as one of the intricate tasks. Therefore, there is a necessity to develop a decision support system to forecast the cardio vascular disease in a patient. Machine learning plays a vital part in disease prediction. In this paper, various machine learning methods were used to predict the heart disease and their performances were compared. The results obtained show the superiority of the Random forest algorithm.

KEY WORDS: CLASSIFICATION ACCURACY, HEART DISEASE, MACHINE LEARNING.

INTRODUCTION

Data mining is used to examines and unearths important information from a massive collection of data. This can be further helpful in exploratory and illustration out patterns for making intelligent business-related decisions. One of the most threatening in medical domain is heart disease, which occurs instantly when its limitations are reached. Machine learning plays a vital role in disease prediction Rajathi N et al,. Cardiovascular disease generally refers to narrowed or blocked blood vessels, which can also lead to heart attack, chest pain or stroke. In general, blood pressure, cholesterol and pulse rate are the main reasons for a heart attack. Heart attack is the main heart disease.

ARTICLE INFORMATION

*Corresponding Author: rajathi.in.it@kct.ac.in
Received 9th Oct 2020 Accepted after revision 7th Dec 2020
Print ISSN: 0974-6455 Online ISSN: 2321-4007 CODEN: BBRCBA

Thomson Reuters ISI Web of Science Clarivate Analytics USA and Crossref Indexed Journal

NAAS Journal Score 2020 (4.31) SJIF: 2020 (7.728)

*Online Contents Available at: http//www.bbrc.in/
Doi: http://dx.doi.org/10.21786/bbrc/13.11/6*

Cardiovascular diseases (CVDs) are the most common explanation for global death. It is estimated that 17.9 million die annually. Heart attacks- once characterized as a part of "the old man's disease" but in this era it can causes for more young people. The heart attack occurs when the coronary arteries become blocked. It causes a serious attack when one or more coronary arteries become blocked.

Bad clinical results would be the doorway in the death of a patient. A computer-based support system can be developed to make a good decision in order to achieve correct and cost-effective treatment. Most of the hospitals maintains their patient data in the form of images, texts and numbers using database systems. This data contains much of the hidden information that has not yet explored, which are useful to make right decisions. Therefore, there is a need to develop an excellent system to help the practitioners to predict heart disease before it occurs. This paper mainly concentrates on the prediction of heart disease considering the past heart disease database records.

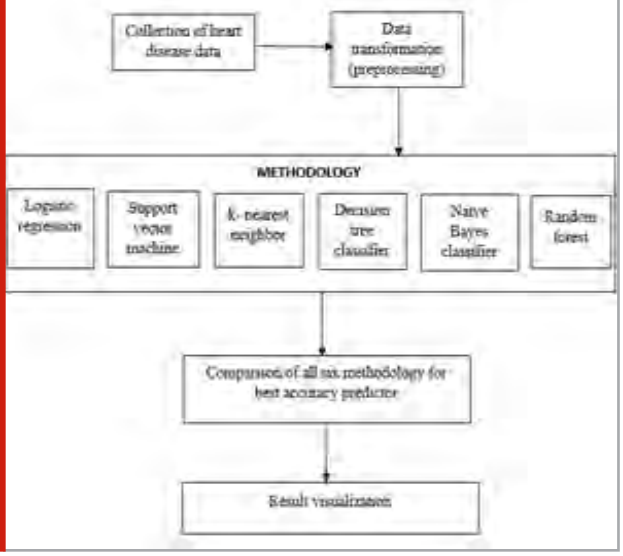
MATERIAL AND METHODS

Various studies with respect to diagnosis of heart disease are discussed in this section. Feixiang Huang et. al., used a data mining process to foresee hypertension from patient medical histories and concluded that J-48 classifier produces better results. M. Amiri et. al., developed diagnosis systems heart sounds. They used 116 heart sound signals to classification and regression trees. M.A. Nishara Banu et. al., used clustering and classification algorithm to forecast the hazard level of the patients. The authors Theresa Princy et. al.., discussed about classification methods including Naïve Bayes, neural network, KNN, decision tree for predicting the risk level of a patient they consider age, gender, pulse rate, blood pressure, cholesterol of each patient.

The various the machine learning algorithms are used by Min Chen et. al., for effective prediction of chronic disease. A multimodal disease risk prediction method was adopted for structured and unstructured data. The prediction accuracy the algorithm is better than other with a convergence speed. Tikotikar A et.al., data mining technique are used in the medical field for clinical diagnosis. It is inferred that an exhaustive survey of medical data help to make well informed diagnosis and decisions.

Cincy Raju et. al.., proved that the SVM technique is an efficient method for predicting heart disease. Praveen Kumar Reddy. M, et. al., used decision tree algorithm to prove the better prediction by comparing its performance with SVM. The authors Akash et. al., applied structured data and the text data of the patient to the k-mean algorithm and archived better accuracy. Reddy et. al., employed machine learning methods for heart disease prediction. All these created an interest to employ machine learning to prediction of heart disease.

Figure 1: Proposed System



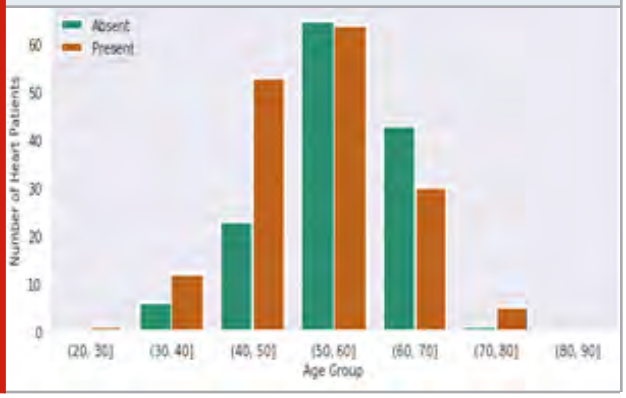
Proposed Methodology: In this paper, various machine learning methods including Naive Bayes classifier, logistic regression, random forest, support vector machine, decision tree classifier and KNN are employed to forecast heart disease. The Python language is used for implementation. The working of the model proposed is pictorially depicted in Figure 1. The dataset is pre-processed in order to remove irrelevant data which helps to achieve better accuracy.

Dataset: The heart disease dataset available in UCI repository taken for this study. The dataset consisting of the parameters including age, sex, chest pain type, serum cholesterol, resting blood pressure etc. After pre-processing the dataset was separated into training (70%) and testing (30%). The models used logistic regression, K-Nearest Neighbor, Support Vector Machine, Naïve Bayes, Decision tree and Random forest are trained using the training data and finally tested with the testing set.

RESULTS AND DISCUSSION

The overall objective of this paper is to forecast more accurately the occurrence of heart disease. Simulation based experiments were conducted using six methodologies named Naive Bayes Classifier, Logistic Regression, Random Forest, SVM, Decision Tree Classifier and KNN. From the result it's been seen that the random forest gives more accuracy as compared as other five techniques. The data set used is decomposed into a training set and testing set. Here, 70% of the dataset is taken for training and the remaining is considered for testing. From the dataset, it is identified that there are more people suffering from heart disease in the 50-60 age group. This is pictorially represented in figure 2.

Figure 2: Number of heart patients in different age group



From the dataset, it is inferred clearly that a greater number of men are suffering from heart diseases as compared to women. While the range of men suffering from heart disease lies between 80-100, the number of women suffering from heart disease lies between 60-80. This is shown in figure 3. The performance of the classification models on the test data was represented using confusion matrix, per class accuracy and classification accuracy and is given in table 1.

Figure 3: Presence of heart disease based on Gender

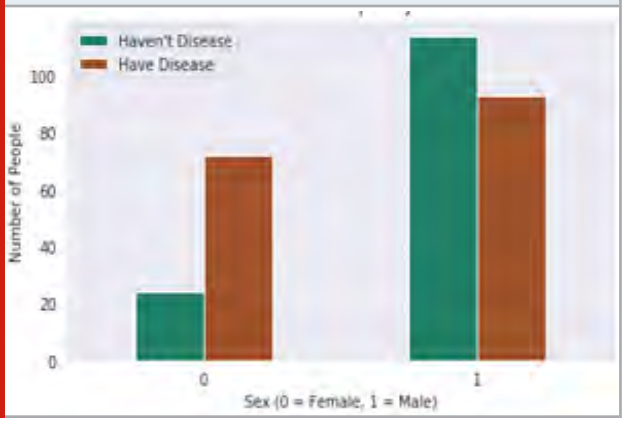


Table 1. Classification Performance of Various Algorithms

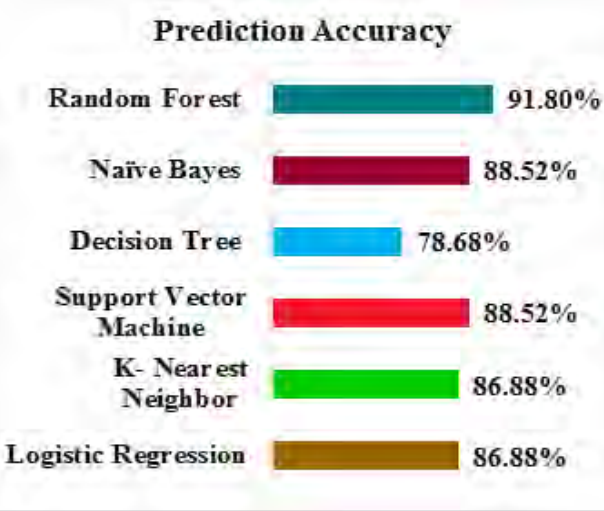
Methodology	Confusion Matrix		Per Class Accuracy
	0 (Female)	1(Male)	
Logistic Regression	23	4	85.18%
	4	30	88.23%
K- Nearest Neighbor	23	4	85.18%
	4	30	88.23%
Support Vector Machine	23	4	85.18%
	3	31	91.17%
Decision Tree	21	6	77.77%
	7	27	79.41%
Naïve Bayes	24	3	88.88%
	4	30	88.23%
Random Forest	25	2	92.59%
	3	31	91.17%

The classification accuracy of various algorithms is graphically represented in figure 4 and the results are presented in table 2. From the results achieved it is inferred that random forest algorithm gives best prediction accuracy than other algorithms.

Table 2. Classification Accuracy of Classifiers

Methodology	Prediction Accuracy
Logistic Regression	86.88%
K- Nearest Neighbor	86.88%
Support Vector Machine	88.52%
Decision Tree	78.68%
Naïve Bayes	88.52%
Random Forest	91.80%

Figure 4: Performance of Classifiers



CONCLUSION

In the field of disease prediction, machine learning plays a significant role. In this paper, various machine learning approaches are used for heart disease forecast. The experimental results illustrate that the Random Forest algorithm achieves the highest accuracy of 91.8% and thus successfully achieving the objective of improving the prediction accuracy. The future work is towards more investigation on evolutionary computation techniques for the problem undertaken and study their performances.

REFERENCES

Amiri, A.M. and Armano, G., 2013, August. Early diagnosis of heart disease using classification and regression trees. In The 2013 International Joint Conference on Neural Networks (IJCNN) (pp. 1-4).

Banu, M.N. and Gomathy, B., 2014, March. Disease forecasting system using data mining methods. In 2014 International conference on intelligent computing applications (pp. 130-133). IEEE.

Chen, M., Hao, Y., Hwang, K., Wang, L. and Wang, L., 2017. Disease prediction by machine learning over big data from healthcare communities. IEEE Access, 5, pp.8869-8879.

Huang, F., Wang, S. and Chan, C.C., 2012, August. Predicting disease by using data mining based on healthcare information system. In 2012 IEEE International Conference on granular computing (pp. 191-194).

Jamgade, A.C. and Zade, S.D., 2019. Disease prediction using machine learning. International Research Journal of Engineering and Technology, 6(5), pp.6937-6938.

Prasad, R., Anjali, P., Adil, S. and Deepa, N., 2019. Heart disease prediction using logistic regression algorithm using machine learning. International journal of Engineering and Advanced Technology, 8, pp.659-662.

Praveen Kumar Reddy, M., Sunil Kumar Reddy, T.,

Balakrishnan, S., Syed Muzamil Basha, & Ravi Kumar Poluru., 2019. Heart Disease Prediction Using Machine Learning Algorithm. International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-10.

Rajathi, N., Kanagaraj, S., Brahmanambika, R. and Manjubarkavi, K., 2018. Early detection of dengue using machine learning algorithms. International Journal of Pure and Applied Mathematics, 118(18), pp.3881-3887.

Raju, C., Philipsy, E., Chacko, S., Suresh, L.P. and Rajan, S.D., 2018, March. A Survey on Predicting Heart Disease using Data Mining Techniques. In 2018 Conference on Emerging Devices and Smart Systems (ICEDSS) (pp. 253-255). IEEE.

Thomas, J. and Princy, R.T., 2016, March. Human heart disease prediction system using data mining techniques. In 2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT) (pp. 1-5). IEEE.

Tikotikar, A., & Kodabagi, M., 2017. A survey on technique for prediction of disease in medical data. In 2017 International Conference on Smart Technologies for Smart Nation (Smart Tech Con) (pp. 550-555). IEEE

Biosc.Biotech.Res.Comm. Special Issue Vol 13 No 11 (2020) Pp-28-33

BBRC
Bioscience Biotechnology Research Communications

Advancement in Identification and Classification Framework for Malaria Parasite Based on Image Manipulation

Alamelu. M*[1] and Kavi Priya. C.U[2]
[1]*Associate Professor, Department of Information Technology, Kumaraguru College of Technology, Coimbatore, India.*
[2]*PG Scholar, Department of Information Technology, Kumaraguru College of Technology, Coimbatore, India.*

ABSTRACT

The transmission of malaria disorder is done by Anopheles genus female mosquitoes. This type of genus mosquitoes is a solitary parasite. The mosquitos that are infected, while sucking the blood it will pass through its salivary glands. The plasmodium then gets injected to the human's blood. In red blood cells the parasites will follow certain division. It gets burst out and it will throw oneself into other RBC's by spreading the parasite. Identifying these type of parasite earlier will reduce the death values across globe. To reduce and identify the parasite in the RBC we proposed an image manipulation - followed totally Malaria disorder discovery framework called the advancement in classification and identification framework for malaria parasite detection. The proposed approach can be analyzed with the data set and categorize the (HSV) histograms, and HSI hue channel histogram using the classification approach and there include the procedure for identification of plasmodium in red blood cells. The Detection and identification of cells based on features algorithm will improve the accuracy of red-blood-cells (RBC) classification using the proposed approach and compare the accuracy with the existing algorithms.

KEY WORDS: ADVANCEMENT IN IDENTIFICATION & CLASSIFICATION FRAMEWORK, MEAN OUTLINE, PARASITE CELL, RED BLOOD CELLS.

INTRODUCTION

Irresistible infections result approximately about more than 30% of deaths globally and basically it emphasized that malaria is one of the three most powerful infections with other harmful inflammation. Agreeing to the later record around 434,000 passing cases had been assessed because of malarial disorder. In the existing method, a conventional microscope is used for malaria parasites. Consumes much time, requires talented labors and the outcome completely depends on the intensive microscopist (Bashar, M. K, 2019). So the detection of stained objects is required for the detection of malaria plasmodium. Moreover, analysation of the stained images for the determination of cells to identify inflamed and noninflamed diagnosis. While detecting the intensity of cells at a fixed and parallel values in image the system can able to detect all the inflamed images (Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G, 2017). There exist certain techniques for malaria disorder. To classify the malaria parasite certain methods are to be used. The methods include acquisitions of image, preprocessing of image, smoothing of image and extension in image segmentation. There include.

ARTICLE INFORMATION

*Corresponding Author: alamelu.m.it@kct.ac.in
Received 15th Oct 2020 Accepted after revision 10th Dec 2020
Print ISSN: 0974-6455 Online ISSN: 2321-4007 CODEN: BBRCBA

Thomson Reuters ISI Web of Science Clarivate Analytics USA and Crossref Indexed Journal

NAAS Journal Score 2020 (4.31) SJIF: 2020 (7.728)

Online Contents Available at: http//www.bbrc.in/
Doi: http://dx.doi.org/10.21786/bbrc/13.11/7

two phases in architectural model (i) Preparation part and (ii) Identification part. In this work we focus on Advancement in classification and identification framework for malaria disorder. (Priyadarshini Adyasha Pattanaik1, Mohit Mittal, 2019), CAD scheme is used to determine nearness of this disorder in RBC images. To identify a presence of malarial disorder using microscopic images this research presents about CAD scheme using deep learning techniques. The outcome of this work gives good result in detecting the parasite and in exactness.

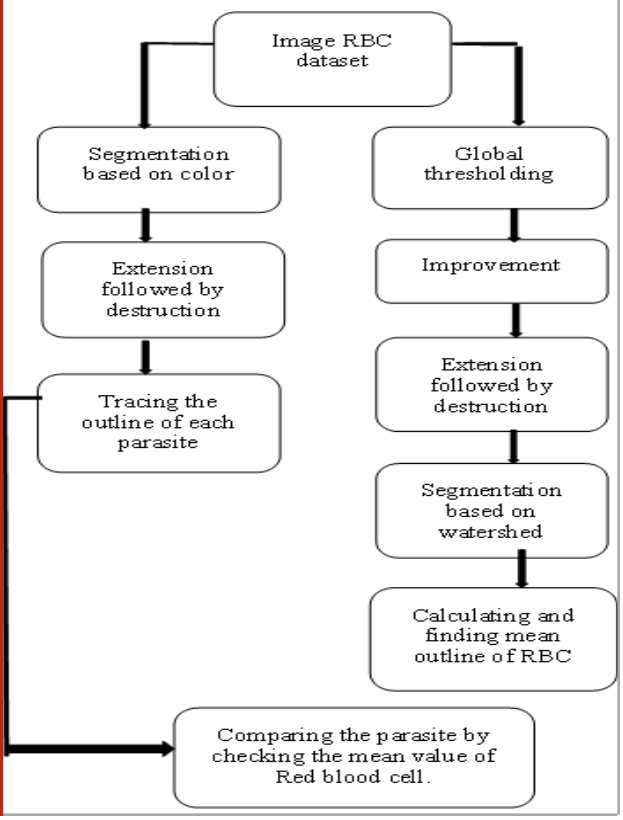
(Karthik, G., Muttan, S., Saravanan, M. P., Seetharaman, R., & Vignesh, V, 2019) provides a novel computerized analysis of malaria via microscopic pictures with the assist of image processing. This community identifies and classifies the crimson blood cells infected by using Plasmodiums like falciparum, vivax, ovale and malariae in skinny blood smears. The image processing set of rules developed is able to locate the infected cells and parasites present in red blood cells. (Bashar, M. K, 2019) examined, a Superintend method to classify malarial disorder stages from microscopy photos has put forward. An analyzation and computational technique has put forward for categorising the lifestyles cycle ranges of the malaria parasite using the some classifiers of support vector machine with numerous features like texture and color. To attain all these benefits, primarily the framework needs to have an excessive computational energy; and furthermore, to make the framework ready, there need to be massive records. There include numerous imperfections which are actually pointed out to distinguish with the tiny investigation.

Related Work: (Yang, F., Poostchi, M., Yu, H., Zhou, Z., Silamut, K., Yu, J., & Antani, S. 2018), this work deals with the malaria plasmodium by using smartphones. The processing pipeline for computerized parasite detection parasite screening and type. An IGMS is used for immediate screening of a whole thick smear picture a custom designed CNN version then Classifies every candidate as either parasite or historical past. (Mustare, N., & Sreelathareddy, V, 2017)it offers by an automatic determination for density of parasite in constant with the microscopic images. This work focuses on a singular approach advanced which can detects all the stages of Plasmodium Vivax and leukocytes. (Khalid, A., Haider, Z., & Khosa, I, 2019), this paper recommended a unique technique to discover the existence of today's malarial disorder in blood images. So that the proposed technique has completed considerably better than the opposite methods with a sensitivity latest 97.60% and specificity present day 95.92%.

(Rollin, G., Lages, J., & Shepelyansky, D. L. 2018), this examine, a aggregate technique include dual thresholding and BLOB examination is proposed for categorising the erythrocytes usage of information and also used ARR for computing cellular place. This research, PPV and Sensitivity about 84.43% and 85.5% in common is performed. (Roy, K., Sharmin, S., Mukta, R. B. M., Sen, A., Roy, K., Sharmin, S., & Sen, A. 2018), the usual technique used to come across malaria parasites in blood is a 'gold standard' traditional method. Wherein professionals discover malaria parasites manually by checking each and every slide .The paper suggests the proper dedication of malaria parasites as conventional requirements might also percolate human mistakes. So the method is proposed that the use of image preprocessing,

photograph segmentation, filtering, class and finally the detection of the malaria parasite. (Dave, I. R. 2017), an automated method involving image processing strategies that are effective of finding and acknowledging the parasite infection inside the images viewed through microscope. Finally the proposed technique produce better accuracy for training, and as a final elegance it produce 60%. The set of rules evolved for class in a hierarchical way confirmed exact outcomes. (Mohammed, H. A., & Abdelrahman, I. A. M. 2017), in this paper the opportunity of the fast as well as correct computerized analysis of RBC disorders that elaborates a model for detecting and classifying this parasites in sampled blood images received through mild microscope. The BFF neural community produced the very best overall performance.

Figure 1: Advancement in Identification & Classification Framework



MATERIAL AND METHODS

Proposed Methodology: The classification framework approach is used to classify and discover malaria parasites as inflamed or uninfected cells. The system

is extreme to differentiate samples among healthful and parasite infected cells. It deals with the work that contains photo reputation and photo classification that calls for a scientific collection of a couple of activities to achieve the goal. The obtained photos undergo a few pre-processing strategies.

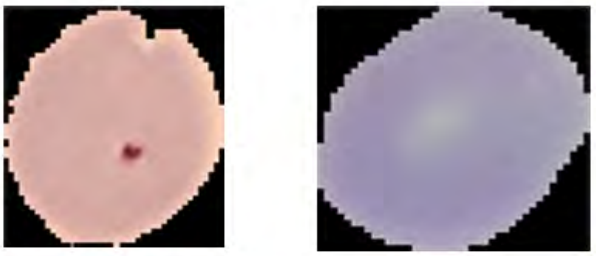
The pre-processing is performed to shape images extra suitable for the subsequent technique. Pre-processing methods contain discount in noise, resizing, and photo differentiation. Normalizing the image size is fundamental for retaining up the spatial resolution of images from exceptional assets. Functions are classified by using the suitable classifier after which it can be set in distinct training. By choosing a parameter suitably which can definitely describe the image Functions are concluded from pix or amounts like signatures and histograms. The procedure utilized for malaria inflamed red blood cells counting is shown in Fig (1).

The steps include,

1. As an input, from the dataset images of red blood cells are taken.
2. Parasite will get segmented based on the color of RBC's.
3. Thresholding is one of best way in segmenting the regions, it will separate the darker and lighter spaces in the cells.
4. The outline of each parasite is traced with that it will calculate the mean outline of RBC.

Red Blood Cell Dataset: The framework chooses the only highlights for making ready by way of disclosing all the separating properties of the approaching picture by means of convolution manner (Bibin, D., Nair, M. S., & Punitha, P.,2017).. Those highlights then skip via the layers in the network.

Figure 2: Cell image



(i) Parasite cell (ii) uninfected cell

Segmentation: An image can be segmented by its essential zone and entity. In Fig 2 images in dataset are named as parasite cell and uninfected cell which shows the difference between them (A. Sai Bharadwaj Reddy and D. Sujitha Juliet, 2019). Exactness of the segmentation explains in the end of successfulness or non-success of procedure analysis. For enhancing the mean value of segmented exactness, a major care must be taken. In many algorithm segmentation similarity, and discontinuity are the general measures of mean values (Roy, K., Sharmin, S., Mukta, R. B. M., Sen, A., Roy, K., Sharmin, S., & Sen, A., 2018). Watershed segmentation is a type that has to be used in the case of implementation. In Fig 1 Watershed segmentation is to be used on the red blood cell images to discrete the object that are in collision.

Segmentation Based On Watershed: Segmentation based on watershed is usually used to discrete the objects that are in collision. Thus in this research watershed segmentation is used so can collided red blood cells can be discrete and helpful in separately counting the red blood cells.

Figure 3: Authentic watershed image

(a) Gradient image (b) Transformed watershed

1. We assumed two dimensional, gray scaled images as homomorphism for which the parameters of images are given by Fig (3)
2. a, b and the intensity of image is corresponded by the heights.
3. Unwanted noises and some small areas within the image will result generally in local remains catchment basins in lower level.
4. For segmentation authentic image dataset can also be used, but using of gradient photo will usually lead along the edges of object high or low. Thus the ridges of watershed can lie beside the edges of an object. By calculating the spaces between closest non-zeros from each picture element, the distance negation is done.

Thresholding: Exactness of the segmentation explains in the end of successfulness or non-success of procedure analysis. For recognizing the mutual plot region segmentation is used. Thresholding is one of best way in segmenting the regions, it will separate the darker and lighter spaces in the cells. Thresholding value is to 0 for the binary photos created by threshold and 1 to every picture element threshold.

Detection and Identification of Cells: To accomplish the end result by classifying the cell images whether they are inflamed by disorder or not is done using feed forward backpropagation neural network (BFFNN). The generalized property will make a framework to train on a corresponding pair of target values and without even training a framework on every possibly applicable pair

of input and output. In spite of predicting the number of infected cells, we focused on identifying the infection type.

Figure 4: Perception and recognition of cells based on features

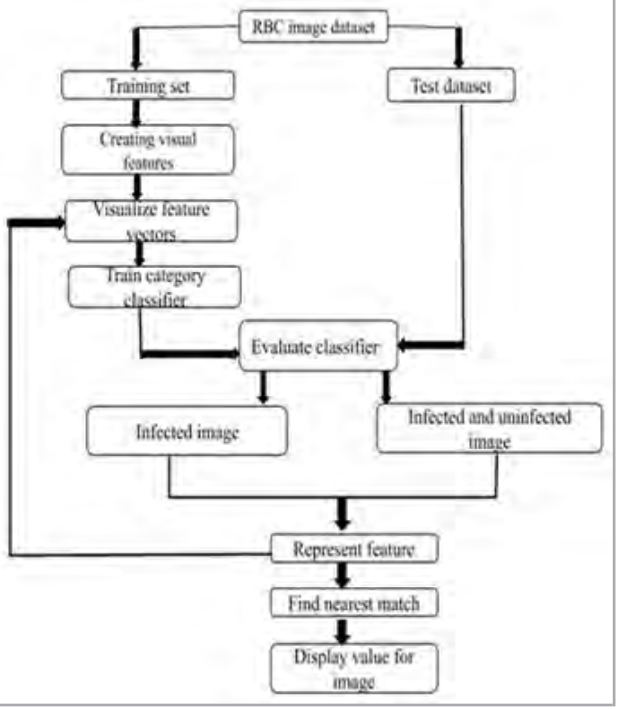
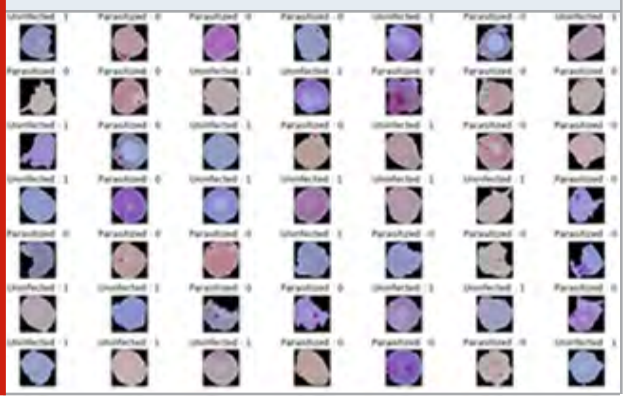


Figure 5: Image database



So algorithm will take the input images as example and generate result by differentiating them automatically. The minimization in the squared errors is the goal of training the data and these processes will get stopped, when there is an increase in validation of error. Loaded input parasite images will pass through a neural network and provide results for each and every initialized weight. Back propagation will help in adjusting all the weights in the framework thus the outcome will come nearer to the known.

In Figure 4. Perception and recognition of cells based on features technique follows the process of extracting the features and displaying the estimated value of image.

Table 1. Evaluation of Comparative Approaches

YEAR	EXISTING RESEARCH ANALYSIS	TECHNIQUES USED	MERITS
2017	Malaria Parasite Detection from Peripheral Blood Smear Images using Deep Belief Networks	Deep neural network (DNN), restricted Boltzmann machine (RBM)	It helps in advancement of machine supported based pattern recognition for malaria plasmodium.
2018	Malaria Parasite Detection And Cell Counting For Human And Mouse Using Thin Blood Smear Microscopy	Support vector machine, Artificial neural network, Multiscale laplacian of gaussian cell detection method	Approximate centers for separate cells are detected and it provides cell detection of accuracy in higher level as well as speed processing.
2018	Image analysis and machine learning for detecting malaria.	Gray Level Run Length Matrix, Quantitative PhaseImaging, SUSAN, Support Vector Machine	It suits malaria patients' test level in larger and allows more standardized corrections, value testing in extensive levels.
2020	Identification of Plasmodium falciparum Stages Using Support Vector Machine Method	Gray Level Co-occurrence Matrix , Otsu threshold method for segmentation	It produced an accuracy of nearly 92% using extracted features combined with the characteristics of morphology.

1. Firstly the specifications are compared with the estimated values of the red blood cell images to check inflamed and non-inflamed cells.
2. But if the value of the cell are equal or less than mean value, it will not do anything.
3. While comparing, when the outside of plasmodium is higher than the mean value of red blood cells, it will highlight it with a circle on all sides of plasmodium

and inflamed cells are get counted.
4. Cells are highlighted by circles if the value of parasites is higher than mean value.
5. If outside of plasmodium lies on or lower than estimated value of the red blood cells it won't do anything.

RESULTS AND DISCUSSION

The dataset taken into consider are of images that are split into train and test data which contain both infected and uninfected cells of nearly 28,000 images. Afterward the Convolution neural network appears is prepared, and the appearance are anticipated. These yields helps in choosing the contaminated as well as sound sampled blood. To maintain uniformity on the images within the dataset pre-processing is required. It makes a difference in preparing the show far off superior by giving more varieties in pictures and by centering on the specified parameters. Reduction of images are done by annotating the images, where the falsely and suspicious images are removed. In Figure. 2 datasets are mentioned as infected (0) cells and uninfected (1) cells. The number of falsely and suspicious images are about 648 removing that from dataset results in 26,164 images. To predict the output these labels are to be used in the model.

Table 1. represents the evaluation of the comparative analysis of the proposed approach with the existing approaches. The proposed approach used an authenticated red blood cell images among nearly 2000 cell images. Identification of malaria disorder is done by using image manipulation, segmentation based on watershed, extension followed by destruction. The testing analysis is done by comparing the separated perimeter cell.

CONCLUSION

In the proposed Advancement in Identification & Classification Framework approach we analyzed the identification of malaria parasites based on image manipulation techniques using few deep learning mechanisms. It will focus on detection of infected red blood cells based on its features. By using this proposed approach both the infected and uninfected red blood cells will obtain better accuracy in comparison with the existing systems. In spite of predicting the number of infected cells, we focused on identifying the infection type. So algorithm will take the input images as example and generate result by differentiating them automatically. The approach can be further expanded by deploying certain techniques and measures in detecting this malaria disorder in earliest.

REFERENCES

Abbas, N., Saba, T., Rehman, A., Mehmood, Z., Kolivand, H., Uddin, M., & Anjum, A. (2019). Plasmodium life cycle stage classification based quantification of malaria parasitaemia in thin blood smears. Microscopy research and technique,IEEE Access, 82(3), 283-295.

Bashar, M. K. (2019, November)" Improved Classification of Malaria Parasite Stages with Support Vector Machine Using Combined Color and Texture Features". In 2019 IEEE Healthcare Innovations and Point of Care Technologies,(HI-POCT) (pp. 135-138). IEEE.

Bibin, D., Nair, M. S., & Punitha, P. (2017). Malaria parasite detection from peripheral blood smear images using deep belief networks. IEEE Access, 5, 9099-9108.

Dave, I. R. (2017, March). Image analysis for malaria parasite detection from microscopic images of thick blood smear. In 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET) (pp. 1303-1307). IEEE.

Ghanmode, B. D., & Paikrao, P. L. (2018, October). "Malaria Parasite Concentration Determination Using Digital Image Processing" In 2018 4th International Conference for Convergence in Technology (I2CT) (pp. 1-4). IEEE.

Gonzalez, R. C., Woods, R. E., & Eddins, S. L. (2004). Digital image processing using MATLAB. Pearson Education India.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications, IEEE Access,vol.73, 220-239.

Hu, C., Ju, R., Shen, Y., Zhou, P., & Li, Q. (2016, May). Clinical decision support for alzheimer's disease based on deep learning and brain networks. In 2016 IEEE International Conference on Communications (ICC) (pp. 1-6). IEEE.

Karthik, G., Muttan, S., Saravanan, M. P., Seetharaman, R., & Vignesh, V. (2019, January). "Automated Malaria Diagnosis Using Microscopic Images". In 2019 Third International Conference on Inventive Systems and Control (ICISC) (pp. 514-517). IEEE.

Khalid, A., Haider, Z., & Khosa, I. (2019, January). Malarial Parasite Detection and Recognition using Microscopic Images. In 2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST) (pp. 304-308). IEEE.

Mohammed, H. A., & Abdelrahman, I. A. M. (2017, January). Detection and classification of malaria in thin blood slide images. In 2017 International Conference on Communication, Control, Computing and Electronics Engineering (ICCCCEE) (pp. 1-5). IEEE.

Mustare, N., & Sreelathareddy, V. (2017, September). :Development of automatic identification and classification system for malaria parasite in thin blood smears based on morphological techniques". In 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI) (pp. 3006-3011). IEEE.

Poostchi, M., Ersoy, I., McMenamin, K., Gordon, E., Palaniappan, N., Pierce, S., & Palaniappan, K. (2018). Malaria parasite detection and cell counting for human and mouse using thin blood smear microscopy. Journal of Medical Imaging,IEEE Access, 5(4), 044506.

Poostchi, M., Silamut, K., Maude, R. J., Jaeger, S., & Thoma, G. (2018). Image analysis and machine learning for detecting malaria. Translational Research, IEEE Access,194, 36-55.

Rajaraman, S., Antani, S. K., Poostchi, M., Silamut, K., Hossain, M. A., Maude, R. J., & Thoma, G. R. (2018). Pre-trained convolutional neural networks feature extractors toward improved malaria parasite detection in thin blood smear images. IEEE Access,vol. 6, 4568.

Ravì, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., & Yang, G. Z. (2016). Deep learning for health informatics. IEEE journal of biomedical and health informatics, IEEE Access,21(1), 4-21.

Rollin, G., Lages, J., & Shepelyansky, D. L. (2019). World Influence of Infectious Diseases From Wikipedia Network Analysis. IEEE Access, 7, 26073-26087.

Roy, K., Sharmin, S., Mukta, R. B. M., Sen, A., Roy, K., Sharmin, S., & Sen, A. (2018). Detection of malaria parasite in giemsa blood sample using image processing. International Journal Of Computer Science & Information Technology (Ijcsit), IEEE.

Setianingrum, A. H., Wardhani, L. K., Ridwan, A. F., & Nasution, S. F. (2019, November). Identification of Plasmodium falciparum Stages Using Support Vector Machine Method. In 2019 7th International Conference on Cyber and IT Service Management (CITSM) (Vol. 7, pp. 1-5). IEEE.

Yang, F., Poostchi, M., Yu, H., Zhou, Z., Silamut, K., Yu, J., & Antani, S. (2019). "Deep learning for smartphone-based malaria parasite detection in thick blood smears". IEEE Journal of Biomedical and Health Informatics, 24(5), 1427-1438.

Biosc.Biotech.Res.Comm. Special Issue Vol 13 No 11 (2020) Pp-34-36

BBRC
Bioscience Biotechnology Research Communications

Breast Cancer Identification Using Logistic Regression

S. Sathyavathi[1], S. Kavitha[2], R. Priyadharshini[3] and A. Harini[4]
[1,2]*Assistant Professor, Department of Information Technology, Kumaraguru College of Technology*
[3,4]*Student, Department of Information Technology, Kumaraguru College of Technology.*

ABSTRACT

Best Cancer is a direct result of uncontrolled chest cell advancement. It happens in females and sometimes in folks. It is the second most compelling motivation for death from harmful development in women around the globe. The likelihood of a woman passing on from chest dangerous development is around 1 of each 38 (around 2.6%). Since 2007, passing rates from chest threatening development have remained steady in women more energetic than 50, anyway have started to diminish in more prepared women. The death rate decreased by 1.3 percent consistently some place in the scope of 2013 and 2017. Dangerous development cells fill either in the lobules or in the chest courses. Lobules are milk-making organs, and lines are channels that pass on the milk to the areola from the organs. In this article, we intend to prescribe an approach to manage the assurance of chest threat subject to a collection of data factors portraying a couple of characteristics of sickness cell. This method uses a model called Logistic Regression for AI. Preliminary revelations show that the backslide model proposed is quantifiably critical and has higher precision.

KEY WORDS: BREAST CANCER, LOBULES, DUCTS, MACHINE LEARNING, LOGISTIC REGRESSION.

INTRODUCTION

Breast Cancer is a disorder in which breast cells develop out of control. There is a couple of breast cancer forms. The type depends on which cells transforms into cancer in breast. Breast cancer accounts for 25 percent of all cancer cases diagnosed in women as a worldwide figure from IARC (Indian Astrobiology Research Center). Around 53 percent of these cases come from developed countries, which account for 82 percent of the global population. Approximately 276,480 new cases of invasive breast cancer are predicted in 2020.Machine Learning has become a critical part of research in medical imaging. Over the years, Machine Learning approaches have progressed from manual seeded inputs to today's automated initialization. As the learning ability of machine learning methods is continuously improving, developments in the field of machine learning have led to more intelligent and self -reliant computer assisted diagnosis (CAD) systems. We have proposed the logistic regression method to predict whether the patient has a malignant or benign tumor based on attribution.

Diagnosis during the early stages of life significantly enhances the future of women with breast cancer, by allowing for therapy as the cancer is rapidly developing. Machine learning is an application that gives systems the capability to learn and develop automatically from knowledge without being specifically programmed. Machine learning offers smart alternatives to the study of large data volumes. Machine Learning can generate precise results and analysis by designing quick and efficient algorithms and data-driven models for real-time data processing. In major field of machine imaging machine learning is widely used techniques in the prediction of the cancer diagnosis.

ARTICLE INFORMATION

**Corresponding Author: Sathyavathi.s.it@kct.ac.in*
Received 10th Oct 2020 Accepted after revision 25th Nov 2020
Print ISSN: 0974-6455 Online ISSN: 2321-4007 CODEN: BBRCBA

Thomson Reuters ISI Web of Science Clarivate Analytics USA and Crossref Indexed Journal

NAAS Journal Score 2020 (4.31) SJIF: 2020 (7.728)

Online Contents Available at: http//www.bbrc.in/
Doi: http://dx.doi.org/10.21786/bbrc/13.11/8

MATERIALS AND METHODS

To conduct a series of experiments ,publicly available breast cancer dataset is used .The following are the steps involved : Loading the dataset, Data preprocessing, Splitting the data into test and train, applying logistic regression to objects , evaluating accuracy.

Dataset collection: The dataset is considered from publicly available Kaggle website. The characteristics are determined from a digitized breast mass image the defines the characteristics of the nuclei of the cells present in the image. There are 569 rows and 33 columns in it. For each cell nucleus, the attributes are ID :number, diagnosis (M1=malignant,B1=benign)and ten real valued features are determined for each nucleus.

Data Pre-Processing: The function .info() from the pandas library is helpful to understand the basic properties of data fed. If there are any missing values in the data set, they can be identified and can be preprocessed before fitting into a model for training and perform testing .Preprocessing of data is an integral step as the quality of information and the valuable information that can be extracted from it directly affects our model to learn. The following unnecessary.

features are dropped:
ID: Which cannot be used for classification Class -Label: Diagnosis column

Unnamed Not applicable features: The pre-processed data after removing the missing values, replacing some default values, and preparing them for the training purpose. The following figure illustrates sample features or attributes the are considered after the preprocessing step and providing the accurate data attributes for the training the model with the actual data for the prediction.

Data Splitting: The data is split into train and test data in the ratio 80:20 to check the performance of the trained model. We used sci-kit learn open source machine learning library and imported "train-test split" from "sklearn_model selection" which splits array or matrices into random train and test subsets .The figure below gives the count of sample train and test data set considered for prediction.

Normalization: Normalization is rescaling of real -valued numeric attributes into 0 and 1 range. Data normalization is implemented in machine learning in training less sensitive to the scale of the features. We have used MinmaxScaler() from the sklearn library for the normalization operation.

Logistic Regression: A popular machine learning algorithm used for classification is logistic regression. It is a statistical model and uses a logistic function to model a binary dependent variable in its basic form. The probability of an observation belonging to a certain class or classification is expected. Logistic regression converts the paradigm of linear regression into classifier and different types of regularization. The most common type of regularization methods are Ridge and Lasso. These two popular methods prevent overfitting. The technique of regularization is used to solve the overfitting issues by penalizing the cost function. The two regularization techniques used for processing are L1 or Lasso regularization and L2 or Ridge regularization.

Hypothesis: Our hypothesis "h1" should satisfy the following condition:

$$0<=h1(x)<=1$$

$$h1(x)=s1(w1t^{1} * x)$$

where x is an observation, s1 is sigmoid function, t1 is time interval and w1 is weights.

COST FOR AN OBSERVATION:
Case 0: h1(x) try to obtain results that are close to 0 as possible

Case 1: h1(x) try to obtain results that are close to 1 as possible

REGULARIZATION
L2 regularization is used for the classification model. The new cost function will be:

$$C(w1) = 1/n \sum^{n}_{i=1} Cost(h(a^{(i)}), b^{(i)}) + lambda/2n \sum nj=1 w1^{2}_{j}$$

The regularization term will heavily control the growth of w1. The h1(x) we obtain with these controlled parameters w1 will be more generalizable. Also, the "lambada "is a hyper-parameter value and found out over cross validation.

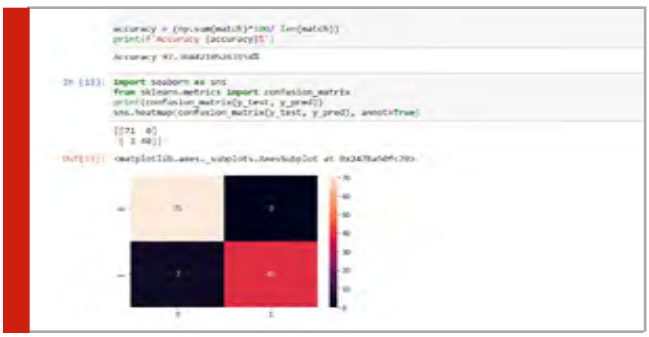
If lambada is greater, it may lead to underfitting If lambada is equal to 0, then there is no regularization effect. Thus, while choosing lambada ,it should be taken care so that the balance for bias vs variance trade -off is balanced properly.

Logistic Regression Parameters: Learning rate: For the advancement calculation (Gradient Descent) ,it is a tuning boundary that characterizes the progression at every cycle while moving towards a least clock work

Max_iter: Maximum number of iterations taken for the optimization algorithm to converge

Penalty: To perform L2 regularization

Tolerance: Value showing the weight between ages in which angle drop to be ended



The subsequent stage is fitting the model as per the training data. The samples are taken to perform probability estimation and class label prediction operation. About 200 iterations are performed on the training data to be trained with logistic regression.

RESULTS

The confusion matrix, also called as error matrix is a particular table structure in the field of statistical classification. The table structure provides the visualization of the performance of the implemented function. Finally, the accuracy is measured and the confusion matrix is plotted using seaborn and sklearn metrics. The result is as follows: Thus, the accuracy obtained is 97.63%.

CONCLUSION

In this work we led a progression of examination based on the machine learning models to improve breast cancer classification for the given data set. We have indicated that logistic regression method has applied on the training dataset shows the promising results. Our model achieves the accuracy of 97.63%. In future work increased data in the data set can be provided and accuracy can be improved.

REFERENCES

Aishwarya Thangaraju Analysis of classification technique for medical data International journal of advanced research trends in engineering and technology vol 5 special issue 12

Bone D et al., (2015) Applying Machine Learning to Facilitate Autism Diagnostics: Pitfalls and Promises, J Autism Dev Disorf 45(5), pp.1121-1136.

Dua D and Graff C (2019) UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, (2016)"A Dataset for Breast Cancer Histopathological , "A Dataset for Breast Cancer Histopathological Image Classification," IEEE Transactions on Biomedical Engineering, vol. 63, no. 7, pp. 1455–1462

Ganesh N Sharma Rahul Various types and management of breast cancer: An overview Journal of advanced pharmaceutical technology and research 1(2) 109-126

Htet Thazin Tike Thein(2015) An approach for Breast cancer diagnosis classification using Neural network, Advanced computing An international Journal 6(1):1-11

Konstantina Kourou (2015) Machine learning applications in cancer prognosis and prediction ,Computational and structural Biotechnology journal Volume 13 pages 8-17

Seigel RL Miller KD Jemal(2016) A cancer statistics ,2016 A cancer journal for clinicians 2016 :66 (1):7-30

Biosc.Biotech.Res.Comm. Special Issue Vol 13 No 11 (2020) Pp-37-41

BBRC
Bioscience Biotechnology Research Communications

Classification of Mushrooms to Detect their Edibility Based on Key Attributes

V. Vanitha[1], M.N. Ahil[2] and N. Rajathi[3]
[1]*Professor, Department of Information Technology, Kumaraguru College of Technology, Coimbatore, India.*
[2]*Student, Department of Information Technology, Kumaraguru College of Technology, Coimbatore, India.*
[3]*Professor, Department of Information Technology, Kumaraguru College of Technology, Coimbatore, India.*

ABSTRACT

Mushroom is found to be one of the best nutritional foods with high proteins, vitamins and minerals. It contains antioxidants that prevent people from heart disease and cancer. Around 45000 species of mushroom are found to be existing in the world-wide. Among these, only some of the mushroom varieties were found to be edible. Some of them are really dangerous to consume. In order to distinguish between the edible and poisonous mushrooms in the mushroom dataset which was obtained from UCI Machine Learning Repository, some data mining techniques are used. Weka is a data mining tool that has various machine learning algorithms which can be used to pre-process, analyse, classify, visualise and predict the given data. Thus in order to select the attributes that helps in the better classification of mushrooms, Wrapper method and Filter method in Weka are used to identify the best attributes for the classification. The attributes 'odor' and 'spore_print_color' were chosen to be the best ones that contributed to the better classification of edible and poisonous mushrooms. After the identification of the key attributes, classification is performed and decision tree is constructed based on those attributes and its Precision, Recall and F-Measure values are analysed.

KEY WORDS: CLASSIFICATION, FILTER METHOD, KEY ATTRIBUTES, WRAPPER METHOD. .

INTRODUCTION

Mushroom is considered to be one of the super food sources of vitamins, minerals and several nutrients. Mushrooms are low in calories, they are free of fat, cholesterol and gluten and the sodium levels in mushroom are found to be low. Thus, these are some of the facts that make mushroom to be one of the healthier foods. Mushroom contains Vitamin B, Potassium, Copper, Selenium, Complex carbohydrates and many more beneficial nutrients. Scientists have discovered that the mushrooms help in preventing the breast cancer. Mushrooms contain antioxidants that help in preventing the body cells from getting affected by chronic diseases (Jiang, J et al, 2010). There are almost 45000 species of mushrooms present in the world out of which only some are edible (Husaini, M et al, 2018).

Thus, it is important to classify the mushrooms as edible and poisonous. For classifying the mushrooms as edible and poisonous, a dataset containing 8124 instances and 22 attributes of mushroom was obtained from UCI Machine Learning Repository (Dua, D et al, 2019). Then the dataset is preprocessed and it is analysed using Weka and a decision tree is also constructed for the dataset using data mining techniques. The data mining tool, WEKA (Waikato Environment for Knowledge Analysis), includes a collection of data mining algorithms and also contains options for data preprocessing, clustering, classification, regression, visualisation.

ARTICLE INFORMATION

*****Corresponding Author: vanitha.v.cse@kct.ac.in
Received 9th Oct 2020 Accepted after revision 12th Dec 2020
Print ISSN: 0974-6455 Online ISSN: 2321-4007 CODEN: BBRCBA

Thomson Reuters ISI Web of Science Clarivate Analytics USA and Crossref Indexed Journal

NAAS Journal Score 2020 (4.31) SJIF: 2020 (7.728)

Online Contents Available at: http//www.bbrc.in/
Doi: http://dx.doi.org/10.21786/bbrc/13.11/9

Data mining is actually a process where it helps in converting a massive amount of data into some useful information (i.e), it helps in predicting the future results with the help of the past acquired results. In (Ismail, S et al, 2018), the paper aims at studying the behavioural features of mushroom which includes surface and shape of the mushroom and mushroom cap's color, features about the gill and stalk of the mushroom, its odor and many other features. Inorder to select the best features, algorithm such as Principal Component Analysis (PCA) is used and for performing the classification, Decision Tree (DT) algorithm is used. The feature, 'odor' was considered to be the highest ranked feature that helped in achieving the high classification accuracy.

In (Pinky, N.J. et al, 2019), Bagging, Boosting and Random Forests are some of the Ensemble methods that have been used to detect whether the mushroom is edible or not. Thus, good results are obtained by using Random Forests for the models that had fixed features for the test sets. In (Eusebi, C et al, 2008), algorithms like unpruned decision tree, voted perceptron algorithm, covering algorithm that generates only correct rules and the nearest neighbor classifier have been used to analyze the mushroom database. Furthermore, (Lavanya, B et al, 2017) suggests that Data mining algorithms such as ID3, CART, and HoeffdingTree (HT) based on decision tree in R studio software environment can be used to find whether the mushroom is edible or not. The paper (M. Senthamilselvi et al, 2018), aims at comparing Weka and Orange tool by analysing Naïve Bayes algorithm which is probability based and decision tree based J48 algorithm in both the tools by using the Mushroom dataset taken from UCI Machine Learning Repository. It is also found that the J48 algorithm was found to produce better results than Naïve Bayes algorithm.

In (Maniraj, V et al, 2015), the classification process is used to categorize whether or not the mushroom is fit for human consumption and clustering is used in the process of identifying the characteristics of the mushroom. Then the association rule is implemented in order to find the best rule so that a decision can be made to check whether the mushroom is edible or not. Thus, it is found that the result obtained by using the decision tree produces the best results of classifying the edible and the poisonous mushrooms. (Wibowo, A et al, 2018) say that mushrooms can be classified into poisonous and edible using machine learning and data mining techniques. The classification algorithms such as Naïve Bayes,

Decision Tree (C4.5) and Support Vector Machine (SVM) have been used for classification and the experiment is performed with the help of Weka. Results imply that the C4.5 algorithm has almost the same accuracy level as that of SVM, but in terms of speed, C4.5 was found to be faster than SVM. Also, Dutta, M in (Verma, S.K. et al, 2018) stated that Artificial Neural Network, Adaptive Neuro Fuzzy inference system and Naïve Bayes are the techniques used for classification to classify the edible and non-edible mushrooms. The performance of ANN and ANFIS are evaluated using accuracy, MAE, kappa statistic. Thus, the performance of ANFIS was found to be more accurate than ANN. ANFIS also had lowest mean absolute error.

In (Mali H. Hakem Alameady*, 2017), Feed-forward Artificial Neural Network (ANN) is used to classify the mushrooms as edible and poisonous. To train the dataset, Multi-Layer Perception is used. It is likewise used to test the data so that it develops a model that helps in predicting the classification. Neural Connection Version 2.0. is a software used for mining the data. This paper aims at explaining Classification, Multi-Layer Preceptor, Back propagation and other mining activities that have been performed on mushroom dataset so that it predict whether the mushroom is fit to consume or not. There are also several datasets that contains the images of the mushroom inspite of the categorical data.

For that in (Ottom et al, 2019), various techniques like Decision Tree, k Nearest Neighbors (KNN), neural network (NN), Support Vector Machines (SVM) are used over a dataset where the dataset contained only the images of the mushroom with background and without background. Thus, an accuracy of 94% is obtained by using KNN with the help of features that are extracted from the images with real dimensions and for the features which has been obtained from the images, an accuracy of 87% is obtained. This paper aims at identifying the key attributes of the mushroom that contributed in detecting the edibility of the mushroom. The dataset which was in nominal format was converted into numerical format using Python. And by using the techniques, Wrapper method and Filter method, it has been found that the attributes "odor" and "spore_print_color" contributed to the better classification of the mushrooms inspite of the other attributes.

MATERIAL AND METHODS

This section is about the methods that have been implemented in the project. The methods included the Data Collection, Pre-processing (Data Transformation), Attribute selection using Wrapper and Filter methods and Decision tree construction using the key attributes.

Data Collection: Since the project is about detecting the edibility of the mushrooms, the dataset containing 8124 instances and 22 attributes of mushroom have been obtained from the UCI Machine Learning Repository (Dua, D et al, 2019). The attributes of the mushroom mentioned in the dataset have been listed in the Figure 1.

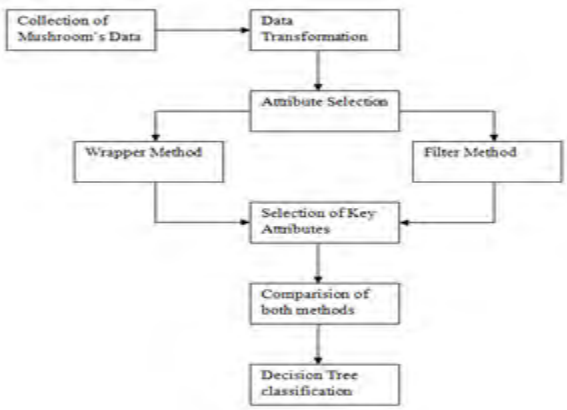
Figure 1: Attributes in the Mushroom Dataset

cap-shape	cap-surface	cap-color	bruises	odor	gill-attachment	gill-spacing	gill-size
gill-color	stalk-shape	stalk-root	stalk-surface-abovering	stalk-surface-belowring	stalk-color-abovering	stalk-color-belowring	veil-type
veil-color	ring-number	ring-type	spore-print-color	Population	habitat		

Pre-Processing: Pre-Processing is an essential process where the data that has been collected may contain null values, missing values, unwanted data that may lead to the production of wrong results. Hence pre-processing is an essential step where in the irrelevant data can be eliminated so there will be a better accuracy in the final results. Weka has an option for preprocessing under the Explorer interface, in which the file that is needed to be preprocessed is uploaded. Then, the necessary steps on pre-processing are performed. With the obtained results from the pre-processing of the dataset, the attributes that are found to be unwanted are removed. The mushroom dataset which has nominal data is encoded using Python. Then, the encoded data was fed into the Weka tool. It was then analysed for any missing values. For instance, in our mushroom dataset, the attribute "Stalk_root" had around 2480 missing values which will obviously does not help in producing the better results. Hence, those kinds of attributes must be removed.

METHODOLOGY

Figure 2: Methodology for classification of Mushroom

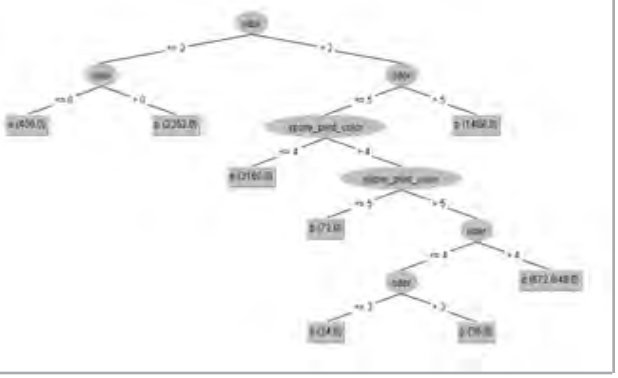


The first step involves uploading the encoded data into the Weka Explorer. Then there is an option called "Select attributes" in the Weka Explorer which has several methods for selecting the highest ranked attributes among a group of attributes. The attribute selection is mainly done to remove the attributes that contributed the least in the classification of mushrooms. It can also be viewed as a method that helps in the dimensionality reduction. Thus in this project, two types of attribute selection have been performed and the results are compared.

Table 1. Key Attributes Selection

KEY ATTRIBUTES SELECTION			
Wrapper Method			Filter Method
Selected Attributes			
Naïve Bayes	J48	Bagging	Ranked Attributes:
'Odor'	'Cap_surface' 'Cap_color' 'Odor' 'Stalk_surface_above_ring' 'Spore_print_color'	'Cap_surface' 'Cap_color' 'Odor' 'Stalk_surface_above_ring' 'Spore_print_color'	'Odor' 'Spore_print_color' 'Gill_color' 'Ring_type' 'Stalk_surface_above_ring'

Figure 3: Decision Tree Representation based on Key Attributes



The first method is the Wrapper method which has a classifier subset evaluator that helps in finding all the possible subsets from the given set of attributes. Then the next step includes applying the classifier algorithm like Naïve Bayes, Bagging and J48. After which a search method must be chosen in order to find the best attributes. The search methods include different types like BestFirst, GreedyStepwise and Ranker. In this project, the classifier algorithms Naïve Bayes, Bagging and J48 and the search method BestFirst were used on the dataset and the key attributes were identified. The second method

that has been used for the attribute selection is the Filter method. The Filter method helps in finding the rank for all the attributes in the dataset out of which one can choose the key attributes based on the rank mentioned for that attribute. In this project, for the filter method, information gain attribute evaluator is used. In order to use the information gain attribute evaluator, rank search method must be used. Then the ranked attributes will appear as an output out of which the top ranked attributes are chosen for classification. The results are tabulated in the below Table I.

Since the attributes "odor" and "spore_print_color" are found to be common in both the attribute selection methods and found to be highly ranked, they are selected as key attributes. Then the selected key attributes are classified under the "Classify" option in the Weka Explorer by using the J48 algorithm and a Decision Tree is constructed based on the key attributes. Its precision, recall and F-measure values are also analysed which were found to be good.

Figure 4: Filter Method Using Ranker

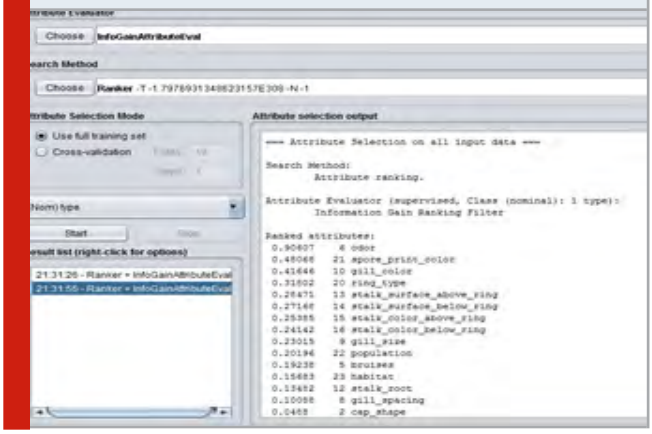
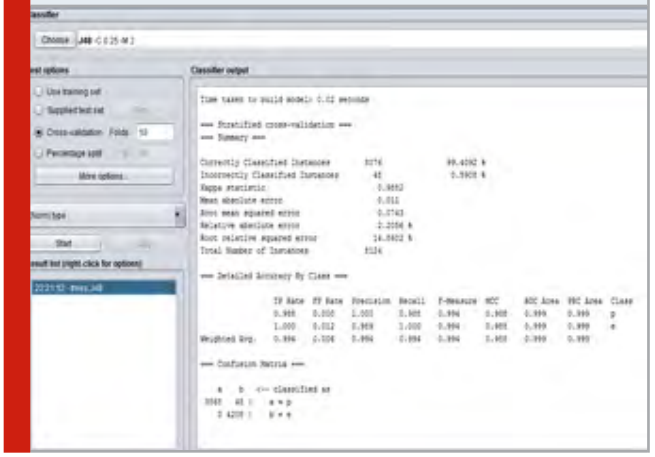


Figure 5: Precision, Recall and F-Measure of Key Attributes



RESULTS AND DISCUSSION

Thus, by using the Wrapper method and Filter method, the Key Attributes that contributed to the better classification of mushrooms are identified. The attributes that have been found to be the best ones from both the attribute selection methods are compared. It is found that both the attribute selection methods almost gave the same results as the output. Hence by using these attributes as the key attributes, there will be a better accuracy in the classification of mushrooms as edible or poisonous.The key attributes were also found to have good Precision, Recall and F-Measure values.

CONCLUSION

This paper discussed about the methods of preprocessing, steps to identify the key attributes that helps in the better classification of edible and poisonous mushrooms and also a comparison between the attribute selection methods in order to find whether both the methods produces the same output.

REFERENCES

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml].

Eusebi, C., Gliga, C., John, D. and Maisonave, A., (2008). Data Mining on Mushroom Database. Journal of CSIS, Pace University, pp.1-9.

Husaini, M., (2018). A Data Mining Based On Ensemble Classifier Classification Approach for Edible Mushroom Identification.

Ismail, S., Zainal, A.R. and Mustapha, A., (2018), April. Behavioural features for mushroom classification. In 2018 IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE) (pp. 412-415). IEEE.

Jiang, J. and Sliva, D., (2010). Novel medicinal mushroom blend suppresses growth and invasiveness of human breast cancer cells. International journal of oncology, 37(6), pp.1529-1536.

Lavanya, B. and Preethi, G.R., (2017) Performance Analysis of Decision Tree Algorithms on Mushroom Dataset. International Journal for Research in Applied Science and Engineering Technology, 5, pp.183-191.

M. Senthamilselvi, P.S.S. Akilashri, (2018). "A Comparative Study on Weka, Orange Tool for Mushroom Data Set", International Journal of Computer Sciences and Engineering, Vol.06, Issue.11, pp.231-236, 2018.

Mali H. Hakem Alameady*, (2017). CLASSIFYING POISONOUS AND EDIBLE MUSHROOMS IN THE AGARICUS AND LEPIOTA FAMILY USING MULTILAYER PERCEPTION. INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY, 6`(1), pp.154–164.

Maniraj, V. and Nithya, J., (2015). Integrating Ontology with Data mining with a Case of Mushroom Analysis. Compusoft, 4(10), p.1983.

Ottom, Mohammad Ashraf. (2019). Classification of Mushroom Fungi Using Machine Learning Techniques. International Journal of Advanced Trends in Computer Science and Engineering. 8. 2378-2385. 10.30534/ijatcse/2019/78852019.

Pinky, N.J., Islam, S.M. and Rafia, S.A., (2019). Classification Edibility Detection of Mushroom Using Ensemble Methods. International Journal of Image, Graphics and Signal Processing, 11(4), p.55.

Verma, S.K. and Dutta, M., (2018). Mushroom classification using ANN and ANFIS algorithm. IOSR Journal of Engineering (IOSRJEN), 8(01), pp.94-100.

Wibowo, A., Rahayu, Y., Riyanto, A. and Hidayatulloh, T., (2018), March. Classification algorithm for edible mushroom identification. In 2018 International Conference on Information and Communications Technology (ICOIACT) (pp. 250-253). IEEE.

Biosc.Biotech.Res.Comm. Special Issue Vol 13 No 11 (2020) Pp-42-46

BBRC
Bioscience Biotechnology Research Communications

Comparative Study of Machine Learning Approaches in Diabetes Prediction

P. Parameswari[1] and N. Rajathi[2]

[1]*Department of Computer Applications, Kumaraguru College of Technology, Coimbatore, India.*

[2]*Department of Information Technology Kumaraguru College of Technology, Coimbatore, India.*

ABSTRACT

Diabetes is a common illness that scares people around the world about their health. Biomedical research effort helps in preventing diabetics and treat it in an efficient way. There are lot of traditional systems, but it cannot handle large amount of data and it leads to problems with high levels of complexity and often it was very tedious. This research helps to design a model that can predict the risk of diabetes in patients with acceptable accuracy. Therefore, to identify diabetes in initial stage, this experiment uses machine learning algorithms, namely Random Forest, J48 as well as Multilayer Perceptron. Experiments are carried out on data collected from the UCI machine learning repository that has been gathered from patients. The impacts of all three algorithms are calculated on many scales, such as Accuracy, Precision, Recall and F-Measure. Accuracy is calculated against instances predicted correctly and incorrectly. The results obtained indicates Random forest performs well with the highest precision of 97.5 percent compared to other algorithms but J48 algorithm took minimum time to build the model.

KEY WORDS: DIABETES, RANDOM FOREST, J48, MULTILAYER PERCEPTRON, MACHINE LEARNING.

INTRODUCTION

Machine learning algorithms can learn from data and evolve from experience, without involving human beings. Learning tasks can include learning the features that map input to output, learning the hidden structure in unlabeled data, also known as example-based learning, where a class label is created from training data by comparing the new instance to memory-stored instances (Goyal et al, 2018). There are three kinds of machine learning algorithms, like Supervised Learning, Unsupervised Learning and Reinforcement Learning. To study a mapping function that converts input variables into output variables, Supervised Learning utilizes labelled training data (Priyanka et al, 2020a & 2020b). Instance -based learning does not create an abstraction from specific instances. Two ways of classification and regression are available for Supervised Learning. Classification is used in the form of categories to predict the result of a given sample data (Parameswari et al, 2015a &2015. b). Unsupervised learning models are used where there are input variables and no corresponding output variables to model the basic structure of the data and are used in unlabeled training data. The reinforcement learning algorithm allows agent to decide the best next step by learning patterns that maximize a reward based on its present state (Sriram et al, 2020).

Diabetes occur when the body is not capable to produce insulin properly, which enables the body to absorb glucose as a cellular fuel and use it. This leads to a persistent increase in blood glucose levels and other abnormalities, leading in turn to the development of complications of the disease. Type I diabetes and type 2 diabetics are the common forms of diabetes. In type

ARTICLE INFORMATION

*Corresponding Author: parameswari.p.mca@kct.ac.in
Received 12th Oct 2020 Accepted after revision 08th Dec 2020
Print ISSN: 0974-6455 Online ISSN: 2321-4007 CODEN: BBRCBA

Thomson Reuters ISI Web of Science Clarivate Analytics USA and Crossref Indexed Journal

NAAS Journal Score 2020 (4.31) SJIF: 2020 (7.728)

Online Contents Available at: http//www.bbrc.in/
Doi: http://dx.doi.org/10.21786/bbrc/13.11/10

I, the immune system initiated a misguided attack and destroyed the pancreas beta cells (Rehman et al, 2020). Approximately 5% of people with diagnosed diabetes are affected and it develops mostly during childhood. The body becomes immune to insulin and destroys the development of insulin in type II diabetes, which affects most organs. It is very important to predict diabetics at an early stage. There are many computer intelligence methods available that help with the available data sets to carry out analysis.

Related Works: The combination of fuzzy logic along with artificial neural networks and case-based reasoning for information engineering techniques has been proposed. Modified co-occurrence and cluster-based mean mode technique was implemented to manage mixed data types that can be used for any type of system (Sokolovska et al, 2018). Optimization of the updated fuzzy ant-miner designed for medical diagnostic efficacy. There are a variety of methods used to produce rules from the medical domain database (Priyanka et al, 2020c). In order to perform a diabetic prediction process, a modified and improved form of artificial bee colony algorithm have been used with an evolutionary algorithm to construct a classification system that allows doctors to make decisions. It is proposed to apply an approach that generates association rules on numeric data to medical data. Predictive Apriori diabetic prediction algorithms have been suggested by generate association rules.

In addition, there is also a greater risk of infection in people with diabetes. In most of the developed nations, diabetes is a leading cause of cardiovascular disease, blindness, kidney failure and lower levels of diabetes (Sierra-Sosa et al, 2019). Development of predictive models in predicting diabetics using risk factors is very significant. Traditional approaches have been proposed by several studies as predictors. Algorithms like Naive Bayes , Logistic Regression and Random Forest assessed the dataset and Random Forest was found to have the highest accuracy on this dataset which we used for this research. Many research come out with prediction of disease by generating rules by using Enhanced Apriori (EA) algorithm with minimum execution time and with better results comparing to the traditional algorithm.

Proposed Methodlogy: For the healthcare field, the Machine Learning Algorithm has tremendous potential as it allows health systems to use medical data for research and improve healthcare at reduced costs. Data mining techniques are useful in the prediction and diagnosis of different health issues, such as heart disease, diabetes, cancer, skin disease and so many more, when applied to health care (Parameswari &tManikantan 2017). A more detailed way in which the information was used to predict diabetics at an early stage was discussed in this part. The proposed system forecasts the diabetics of a individual based on questions and answers provided to the prediction system. The impacts of all three algorithms are calculated on many scales, like Precision, Accuracy, F-Measure, and Recall. Accuracy is calculated against correct and incorrectly predicted instances (Priyanka &tThangavel, 2020).

Random Forest Algorithm: Random Forest is a supervised learning algorithm used for both classification and regression. But it is, however, primarily used for problems with classification. As inferred, a forest is made up of trees, and more trees make the forest stronger. Similarly, the random forest algorithm produces decision trees on data samples and then gets the prediction from every one of them and picks the best solution by polling. It is an ensemble strategy that is stronger than a single decision (Parameswari et al, 2015c).

1. Begin by selecting random samples from a specific dataset.
2. For every sample, this algorithm will create a decision tree. Then, from any decision tree, it will get the prediction result.
3. For each predicted outcome, voting will be carried out at this point.
4. Eventually, as the final prediction outcome, pick the most voted prediction result.

J48 Algorithm: In order to construct a trimmed C4.5 decision tree, Quinlan's C4.5 algorithm updates J48. All aspects of the data are divided into minor subsets based on a decision. J48 looks at the structured data that really results in the information being broken by selecting an attribute. The attributes are used to obtain extremely structured knowledge. By the algorithm, the minor subsets are returned (Nagata et al, 2018). When a subset has a position with a similar class, the split strategies end.

It functions as follows.

1. If the instances belong to a similar class, the leaf is labelled with a similar class.
2. The possible data will be measured for each attribute, and the benefit in the data will be taken from the attribute test.
3. Eventually, based on the current selection criteria, the best attribute will be picked.

Multilayer Perceptron: In Artificial Neural Network (ANN), feed forward neural network class is the Multilayer Perceptron (MLP). The term MLP is applied ambiguously, for any feed-forward ANN, often specifically to mention networks made up of multiple perceptron layers. Multilayer Perceptron is often referred to colloquially as vanilla neural networks, particularly when they have a single hidden layer. The MLP consists of three main layers that is, input layer, hidden layer, and the output

layer (Mir and Dhagae 2018). Every node is a neuron that utilizes a nonlinear activation function, except for the input nodes. MLP uses a supervised learning approach to back-propagation training. MLP is distinguished from a linear perceptron by its several layers and non-linear activation. It can discern data which cannot be separated linearly.

RESULTS AND DISCUSSION

The Weka tool is used in this research work to complete the experiment. Weka is a software tool that includes a collection of different machine learning methods for data classification, clustering, regression and visualization. One of the key advantages of using Weka is that it can be customized to suit the requirements. The main objective of this research is to predict diabetes-affected patients. We have used the algorithm J48 with modified Weka. Updated Weka offers various types of data file classification test choices, such as user training set, test set given and cross-validation. The 10-fold cross-validation data is chosen.

Table 1. Data Set Description

S.No	Attribute	Description
1.	Age	Patient Age
2.	Gender	Male/Female
3.	Polyuria	Sign of diabetics
4.	Polydipsia	Increase in Thirst
5.	Weight loss	Patient having sudden weight loss
6.	Weakness	Happens if cells don't have an adequate amount of glucose
7.	Polyphagia	Extreme hunger
8.	Genital thrush	Yeast infection
9.	Visual blurring	Temporary blurred vision
10.	Itching	Skin itching
11.	Irritability	Disturbance in patient mood.
12.	Delayed healing	Delay in wound healing
13.	Partial paresis	Weakness in muscles
14.	Muscle stiffness	Feeling tightness in muscles.
15.	Alopecia	Disorder that is the reason for hair fall.
16.	Obesity	Unnecessary body fat

Data Set Description: Data was obtained using direct surveys from patients at Sylhet Diabetes Hospital in Sylhet, Bangladesh, donated to the UCI repository. The suggested approach that is taken from the UCI Repository is evaluated. This dataset contains medical descriptions of male and female patients in 520 instances. In the dataset, there are 16 attributes where the value of one class 0 is viewed as diabetes negative and the value of another class 1 is viewed as diabetes positive. Dataset in this research work has 520 instances and sixteen attributes namely and all the 16 attributes are described in the Table 1.

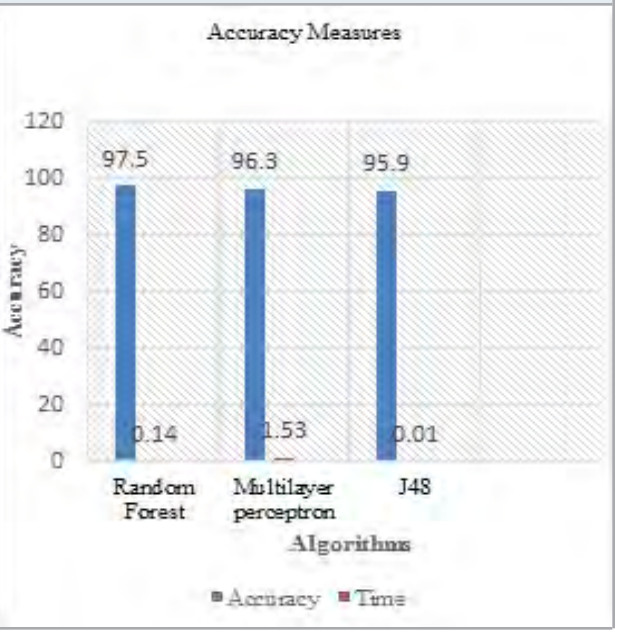
Table 2. Accuracy Measures

Algorithm	Correct Predictions	Incorrect Predictions
Random Forest	507	13
Multilayer perceptron	501	19
J48	499	21

Table 3. Prediction Results

Algorithm	Accuracy	Time (Model Building)
Random Forest	97.5	0.14
Multilayer perceptron	96.3	0.78
J48	95.9	0.01

Figure 1: Accuracy Measures of different algorithm in manipulations



In Figures 1, 2 and 3, the pictorial representations of the prediction results and the time taken to construct the computational model of the three algorithms were given and the results are discussed in Table 2, 3, and 4. Where the difference in time between all three algorithms can be seen. The impacts of all three algorithms were evaluated on different scales such as Precision, Accuracy,

F-Measure, and Recall. Precision is measured against exact and wrongly calculated events. The results obtained indicate that Random Forest performs well with the highest accuracy of 97.5 percent compared to other algorithms, but the J48 algorithm took limited time to create the model.

Figure 2: Prediction Results of different algorithms during manipulations

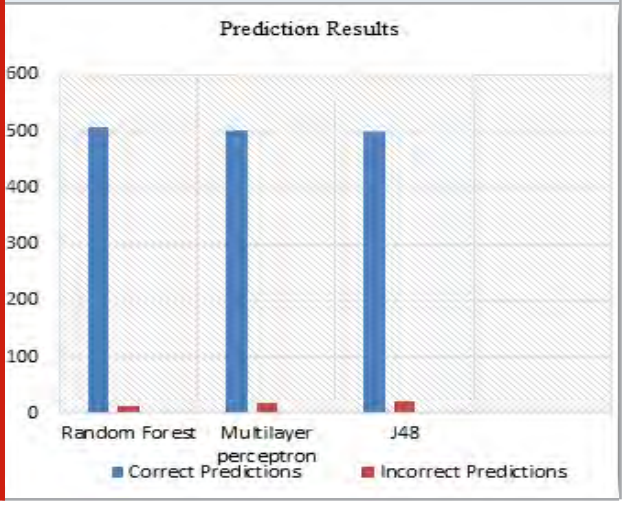
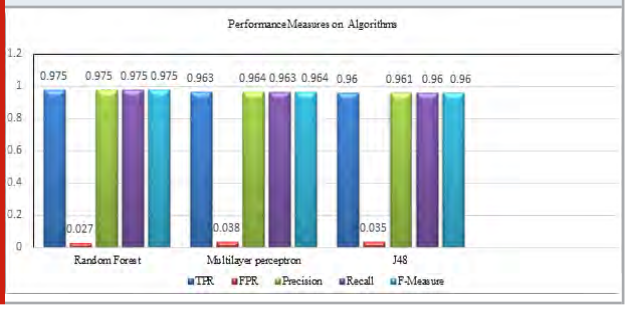


Table 4. Performance of Machine Learning Algorithms for Different Measures

Algorithm Random	TPR	FPR	Precision	Recall	F-Measure
Forest	0.975	0.027	0.975	0.975	0.975
Multilayer perceptron	0.963	0.038	0.964	0.963	0.964
J48	0.960	0.035	0.961	0.960	0.960

Figure 3: Performance of Machine Learning Algorithms for Different Measures



CONCLUSION

Any intelligence systems for the management of diabetics should be established with regard to the high incidence of diabetics and their impact on mortality. So the suggested system may support individuals with their symptoms to predict diabetics at an early stage. This research results in disease prediction through the use of machine learning algorithms that provide diabetic patients with knowledge. The proposed method shows that the impacts of all three algorithms are calculated by different measures such as Precision, Accuracy, F-Measure, and Recall. Compared to other algorithms, the results obtained indicate that Random forest performs well with the highest accuracy of 97.5%, but the J48 algorithm took limited time to construct the model.

REFERENCES

Goyal, M., Reeves, N.D., Davison, A.K., Rajbhandari, S., Spragg, J. and Yap, M.H., 2018. Dfunet: Convolutional neural networks for diabetic foot ulcer classification. IEEE Transactions on Emerging Topics in Computational Intelligence.

Mir, A. and Dhage, S.N., 2018, August. Diabetes disease prediction using machine learning on big data of healthcare. In 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) (pp. 1-6). IEEE.

Nagata, M., Takai, K., Yasuda, K., Heracleous, P. and Yoneyama, A., 2018, July. Prediction models for risk of type-2 diabetes using health claims. In Proceedings of the BioNLP 2018 workshop (pp. 172-176).

Parameswari, P. and Manikantan, D.M., 2017. Geo-Intelligence System: A Frame work for agricultural improvements. International Journal of Pure and Applied Mathematics, 116(12), pp.117-125.

Parameswari, P. and Samath, J.A., 2015a. QOS Based Web Service Ranking Using Fuzzy C-means Clusters. Research Journal of Applied Sciences, Engineering and Technology, 10(9), pp.1045-1050.

Parameswari, P., Abdul Samath, J. and Saranya, S., 2015c. Scalable clustering using rank based preprocessing technique for mixed data sets using enhanced rock algorithm. African Journal of Basic & Applied Sciences, 7(3), pp.129-136.

Parameswari, P., Samath, J.A. and Saranya, P., 2015b. Efficient birch clustering algorithm for categorical and numerical data using modified co-occurrence method. Int. J. Appl. Eng. Res, 10(11), pp.27661-27673.

Priyanka, E.B. and Thangavel, S., 2020. Influence of Internet of Things (IoT) In Association of Data Mining Towards the Development Smart Cities-A Review Analysis. Journal of Engineering Science & Technology Review, 13(4).

Priyanka, E.B., Thangavel, S. and Gao, X.Z., 2020b. Review analysis on cloud computing based smart grid technology in the oil pipeline sensor network system. Petroleum Research.

Priyanka, E.B., Thangavel, S. and Pratheep, V.G., 2020a.

Enhanced Digital Synthesized Phase Locked Loop with High Frequency Compensation and Clock Generation. Sensing and Imaging, 21(1), pp.1-12.

Priyanka, E.B., Thangavel, S., Madhuvishal, V., Tharun, S., Raagul, K.V. and Krishnan, C.S., 2020c. Application of Integrated IoT Framework to Water Pipeline Transportation System in Smart Cities. In Intelligence in Big Data Technologies—Beyond the Hype (pp. 571-579). Springer, Singapore.

Rehman, A., Athar, A., Khan, M.A., Abbas, S., Fatima, A. and Saeed, A., 2020. Modelling, simulation, and optimization of diabetes type II prediction using deep extreme learning machine. Journal of Ambient Intelligence and Smart Environments, (Preprint), pp.1-14.

Sierra-Sosa, D., Garcia-Zapirain, B., Castillo, C., Oleagordia, I., Nuño-Solinis, R., Urtaran-Laresgoiti, M. and Elmaghraby, A., 2019. Scalable healthcare assessment for diabetic patients using deep learning on multiple GPUs. IEEE Transactions on Industrial Informatics, 15(10), pp.5682-5689.

Sokolovska, N., Chevaleyre, Y. and Zucker, J.D., 2018, March. A provable algorithm for learning interpretable scoring systems. In International Conference on Artificial Intelligence and Statistics (pp. 566-574).

Sriram, R.D. and Reddy, S.S.K., 2020. Artificial intelligence and digital tools: future of diabetes care. Clinics in Geriatric Medicine, 36(3), pp.513-525.

Biosc.Biotech.Res.Comm. Special Issue Vol 13 No 11 (2020) Pp-47-51

BBRC
Bioscience Biotechnology Research Communications

In Silico Screening of Antimicrobial Compounds Using Docked Complexes of Antibiotics and Antimicrobial Peptides

Dinakari Sarangan, Keerthana Sakthivadivelan, Darsini Thiyagarajan, Apsara Sudhakar, Krithika Balakrishnan, Ram Kothandan, and Kumaravel Kandaswamy*
[1]*Department of Biotechnology, Kumaraguru College of Technology, Coimbatore - 641049, Tamil Nadu, India.*

ABSTRACT

Biofilms are sessile aggregates of bacterial cells enclosed by a slimy matrix that protect the cells from bactericidal molecules. Biofilm associated infections such as Urinary Tract Infections (UTI) are caused by bacterial strains such as *Escherichia coli* and *Enterococcus faecalis*. Biofilm often exhibits increased resistance to the antimicrobial compounds due to their polymicrobial nature. The matrix of biofilm consists of exopolysaccharides, extracellular DNA (eDNA), and proteins that are crosslinked to provide structural integrity to the biofilms. The proteins in the biofilm matrix are regarded as the potential targets for the antibiotics and the antimicrobial peptides, which kills the bacterial population in the biofilm by disrupting them. Studies have reported that the metabolically active cells in the biofilms can be killed by antimicrobial peptides while the cells with low metabolic activity can be destroyed by antibiotics. In this study, we have used several combinations of antibiotics and antimicrobial peptides, we have obtained a docked complex of Human Beta Defensin 3 (Positively charged peptide) with Ciprofloxacin (Negatively charged antibiotic) and Dermcidin (Negatively charged peptide) with Tobramycin (Positively charged antibiotic). The efficient pair of antimicrobial peptide and antibiotic was then used to dock with biofilm matrix proteins. In essence, this study aims to provide a combinatorial approach to identify drug targets in biofilm associated infections

KEY WORDS: IN-SILICO DOCKING, ANTIMICROBIAL PEPTIDE, BIOFILMS, AND ANTIBIOTICS.

INTRODUCTION

Pathogenic strains such as *Escherichia coli* (*E.coli*) and *Enterococcus faecalis* (*E.faecalis*)are the major cause of Urinary Tract Infection (UTI) and other biofilm associated infections(Madrazo et al., 2020)(Govindarajan et al., 2020) In addition, chronic infections such as cystic fibrosis and periodontitis were also proven to be biofilms associated infections. In order to establish the infection, the pathogenic bacteria need to attach to the host cells. This host- pathogen interaction leads to primary attachment of bacterial pilus to the host surface and aids in colonizing the host epithelium. Pilus of the bacteria are long filamentous proteins extending from bacterial surfaces. These pilin proteins are the contributory factors for many diseases such as cystitis, meningitis, sepsis, porynephritis and UTI(Sillanpää et al., 2010).

The pilus assembly of gram positive and gram-negative bacteria are very distinct. There are five different types of pilus assembly pathway in gram negative bacteria and those are Chaperone–Usher (CU) pili, type IV pili, type IV secretion pili, type V pili and curli fibres(Guillermo

ARTICLE INFORMATION

**Corresponding Author: kumaravel.k.bt@kct.ac.in*
Received 05th Oct 2020 Accepted after revision 07th Dec 2020
Print ISSN: 0974-6455 Online ISSN: 2321-4007 CODEN: BBRCBA

Thomson Reuters ISI Web of Science Clarivate Analytics USA and Crossref Indexed Journal

NAAS Journal Score 2020 (4.31) SJIF: 2020 (7.728)

Online Contents Available at: http//www.bbrc.in/
Doi: http://dx.doi.org/10.21786/bbrc/13.11/11

Garcia-Manero Shao-Qing Kuang, Susan O'Brien, Deborah Thomas, and Hagop Kantarjian, 2005).However, among those pathways CU pili is the most extensively studied pathway. However, in gram positive bacteria, there are only two pathways, one being the well-studied sortase pathway (Telford et al., 2006) and the other the type IV mechanism(Muschiol et al., 2019). Pili in both gram positive and gram negative bacteria is made of major and minor protein subunits. The major pilin subunit is repetitive and more abundant when compared to minor pilin subunits(Giltner, Nguyen and Burrows, 2012).

The Major pilin subunit of CU pili of gram-negative bacteria is fimA and minor pilin subunits are a periplasmic chaperone(fimC), usher (fimD), and a tip adhesion(fimH)(Busch, Phan and Waksman, 2015). In gram positive sortase assembled pili, the major pilin is EbpC and the minor pilin is adhesion pilin EbpA(La Rosa et al., 2016). Therefore, pili proteins are considered as an attractive target for antimicrobial therapy. Studies in the past demonstrated that Antimicrobial peptides (AMPs) such as human Beta Defensin 3 (hBD3) can focally target sortases and its pili proteins(Kandaswamy et al., 2013). Majority of antibiotics such as ampicillin, tetracycline, streptomycin have been used to treat a wide range of bacterial infections but over a period of time bacterial strains have gained resistance to those antibiotics. Therefore, to overcome this, Antimicrobial peptides (AMPs) were first discovered in the early 1980s and AMPs such Human Beta Defensin 5 (hBD5) were proven to kill bacteria (Chileveru et al., 2015).

In the recent years, AMPs such as dermcidin were also proved to be act against pathogens(Schittek et al., 2001), however the bacterial strains acquired resistance to those AMPs making it challenging to treat bacterial infections(Schmidtchen et al., 2002).Therefore, in this study, we have used several combinations of antibiotics and antimicrobial peptides. We have obtained a docked complex of Human Beta Defensin 3 (Positively charged peptide) with Ciprofloxacin (Negatively charged antibiotic) and Dermcidin (Negatively charged peptide) with Tobramycin (Positively charged antibiotic). Furthermore, this study also demonstrate that the biofilm associated pili protein (FimA) can be targeted using docked complexes of AMPs and antibiotics. (yen and Burrows, 2012) The Major pilin subunit of CU pili of gram-negative bacteria is fimA and minor pilin subunits are a periplasmic chaperone(fimC), usher (fimD), and a tip adhesion(fimH)(Busch, Phan and Waksman, 2015).

MATERIAL AND METHODS

Target Selection: The target selection was performed as mentioned in previous studies (Table:1). We have chosen few well studied antibiotics and AMP for docking as mentioned in Table 1.

Retrieval and Preparation of target protein: The crystallized structure of the antimicrobial proteins were retrieved from Protein Data Bank (PDB) andthe energy minimization of proteins was performed using GROMACS(Lemkul, 2019). Then protein was prepared using the protein preparation as mentioned in the previous studies(Madhavi Sastry et al., 2013) the auto dock software assigns missing bonds, bond order, flexible torsions and charges to the input structures during the preparation process and makes them readily available for docking studies(Sivaramakrishnan et al., 2019)

Retrieval and Preparation of ligands: The well-studied antibiotics (as shown in Table 1) and its 3-Dimensional structure was retrieved from the Drug bank and prepared for docking studies. An autodock user module 4.2 was used in this study. The auto dock software assigns missing charges, bonds, bond order and hybridization, detects flexible torsions, creates explicit hydrogens and finally energy-minimized structure can be obtained(Sivaramakrishnan et al., 2019).

Molecular Docking: A blind docking was performed using autodock vina as described in the previous study (Sivaramakrishnan et al., 2019). Molecular docking was performed to understand the interaction of selected AMP's with antibiotics . The Initial docking analysis was performed using the autodock 4.2 package (Sivaramakrishnan et al., 2019). The surface module of autodock creates a double colored molecular surface according to the electrostatic property of the receptor protein. The cavity prediction algorithm predicts the cavities present in the receptor protein and displays it to the user in green color and finds the potential binding sites of the receptor protein. The parameters were set to a molecular surface with extended Van der Waals and number of cavities to five.

The docking was carried out using autoDock simplex evolution search algorithm with grid resolution 30 Å for grid generation and cavity predicted using a search algorithm called cavity prediction algorithm.43 In cavity prediction wizard the number of cavities was restricted to three and the cavity with the large volume was selected as the origin for the binding site. The docking wizard runs with default parameters autoDock as a search algorithm, number of runs, maximum population and maximum iteration was limited to 10, 50 and 1500 respectively. The selected phytochemicals were docked against the receptor proteins and best-generated poses were selected based on the docking scores. The Interaction between the ligand and the receptor protein depends on the number of H-bonds, distance and binding energy. Some poses have favorable hydrogen bond interactions with active site amino acid residues of target bacterial membrane proteins.(Sivaramakrishnan et al., 2019).

RESULTS AND DISCUSSION

Combinational therapy is a promising approach to overcome and mitigate antimicrobial resistance. In combinational therapy, a combination of conventional antibiotics is used together with other antimicrobial peptides to increase the treatment efficacy (Thappeta et al., 2020). Combinational therapy can extend the

lifetime of drugs, inhibits after effects and mitigates the emergence of resistance. While there have been several reports of synergy between conventional antibiotics and other drugs, very few have examined synthetic antimicrobial peptides in combination with conventional antibiotics(Thappeta et al., 2020). In this study, we have docked several antimicrobial peptides and antibiotics to obtain a docked complex of opposite charges (Figure 1 & Table 1) using auto dock vina. The docking score represents the affinity of the antibiotics towards the antimicrobial peptide. More negative the docking score, better the binding affinity.

Table1. Proteins and AMP's chosen for docking

Peptides	PDB ID	Charge	Ligand	Drug bank ID	Charge	References
HBD-3	1KJ6	Positive	Ciprofloxacin	DB00537	Negative	(Dhople, Krukemeyer and Ramamoorthy, 2006) (Walters et al., 2003)
Dermcidin	2YMK	Negative	Tobramycin	DB00684	Positive	(Schittek et al., 2001) (Walters et al., 2003)
Hevein	1Q9B	Negative	Streptomycin	DB01082	Positive	(Prabhu et al., 2013)(Tseng, Bryan and Van den Elzen, 1972)
LL 37	2K6O	Positive	Tetracycline	DB00759	Negative	(Overhage et al., 2008)(Pamp et al., 2008)

Table 2. Estimation of docking scores using autodock vina

Protein and Ligand Complex	Docking score
Human Beta Defensin 3 with Ciprofloxacin	-5.0
Dermcidin with Tobramycin	-5.2
Hevein with Streptomycin	-5.2
LL-37 with Tetracycline	-5.6

Table 3. Scores of FimA docked with Antibiotics

Antibiotics	Scores
ampicillin	-5.6
ciprofloxacin	-5.7
streptomycin	-5.6
tobramycin	-5.2
Tetracycline	-5.9

Table 4. Docking score of protein-protein docking complexes

s.no	Protein complexes (pilin protein-AMPs)	z-score
1	FimA-Havein	-1.4
2	FimA-LL37	-1.8
3	FimA-HBD3	-2.1

Figure 1: Docked image of (a)Ciprofloxacin and Human beta defensin 3 (b)Tobramycin-and Dermcidin (c)Hevein and Streptomycin (d)LL-37 and Tetracycline

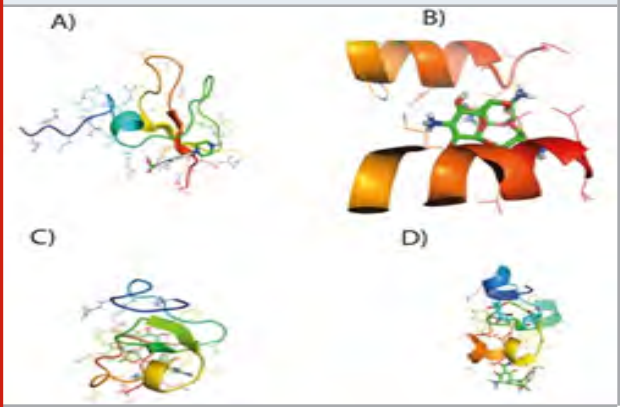
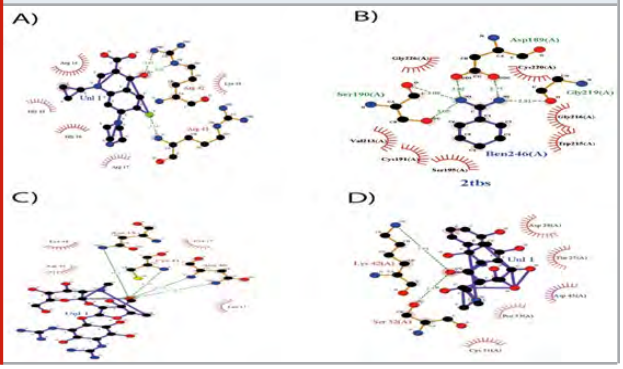


Figure 2: Ligplot image of docked complexes of (a) Ciprofloxacin and Human beta defensin 3 (b)Tobramycin and Dermcidin (c) Hevein and Streptomycin (d) LL-37 and Tetracycline



The complex Dermcidin and tobramycin with an affinity of-5.2(Table 2) has the highest affinity as Tobramycin is docked with ASP 42, ASP 45 ,SER 46 which creates an ionic interaction. Also the shorter distance (< 3 Å) between the peptide and the antibiotic can be clearly seen in the ligplot result (Figure 2 b).From the affinity scores of FimA and antibiotics complexes (as mentioned in Table 3) it is evident that fimA has a higher affinity of -5.9 for tetracycline which can be seen in the ligplot results (Figure 3e) . A 3D image of FimA and antibiotics with

hydrogen bonds can be seen in figure 4. The antibiotic ampicillin form hydrogen bonds with TYR 158, LYS68 (of FimA), Ciprofloxacin form hydrogen bonds with TYR158, SER67, LYS68 , streptomycin form hydrogen bonds with ASP29, GLY26, GLY53, ASN55, THR31, tobramycin form hydrogen bonds with GLN33, THR31, GLN30, GLY26, SER27, ASN 55, and tetracycline form hydrogen bonds with GLN98, THR9 of FimA can be seen in ligplot results (Figure 3).

Figure 3: Ligplot image of docked complexes a) fimA and ampicillin b) fimA and ciprofloxacin c) fimA and streptomycin d) fimA and Tobramycin e) fimA and tetracycline.

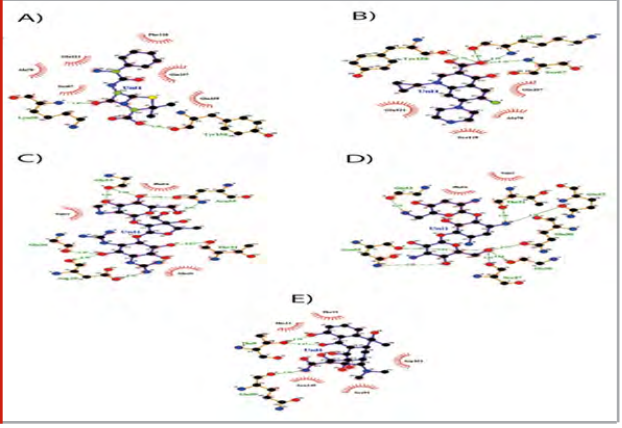
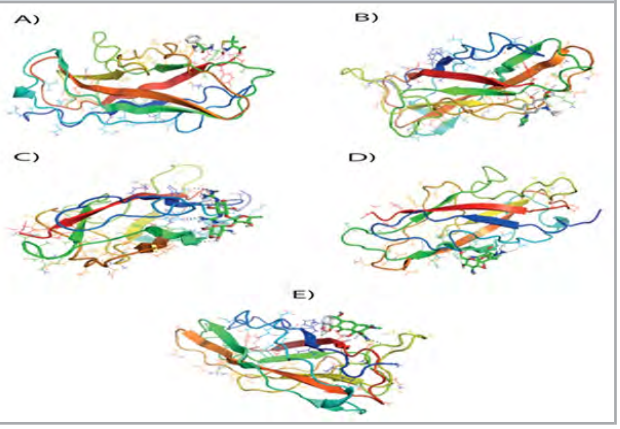


Figure 4: Docked images of fimA with antibiotics a) fimA-ampicillin b) fimA-ciprofloxacin c) fimA-streptomycin d) fimA-Tobramycin e) fimA- tetracycline



Protein - protein Docking was carried out using HADDOCK online tool (Van Zundert et al., 2016) Based on Z score the best docked complex was chosen among clusters. The best docked complexes were chosen for all the antimicrobial peptides with FimA listed in the Table 4. Dimplot and PIC: Protein Interactions Calculator(Tina, Bhadra and Srinivasan, 2007) were used to analyze the interactions between FimA and antimicrobial peptides. The Important interactions between FimA and AMPs based on the bond length are ASP62(A):TYR9(B),ALA 25(A):ARG36(B), VAL123(A):THR35(B) of Human beta defensin(B), LYS155(A):GLN29(B) of hevein,andGLU15

Figure 5: Docked of fimAwith Anti-microbial peptides a) fimA-Hevein b) fimA- Human beta defensin c) fimA-LL-37

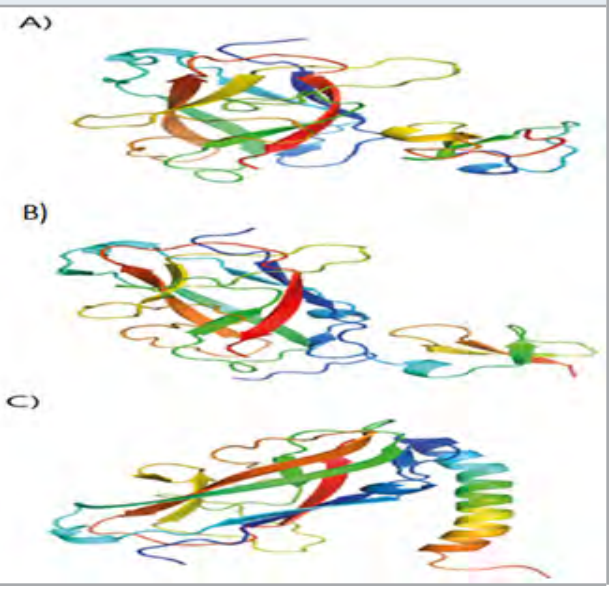
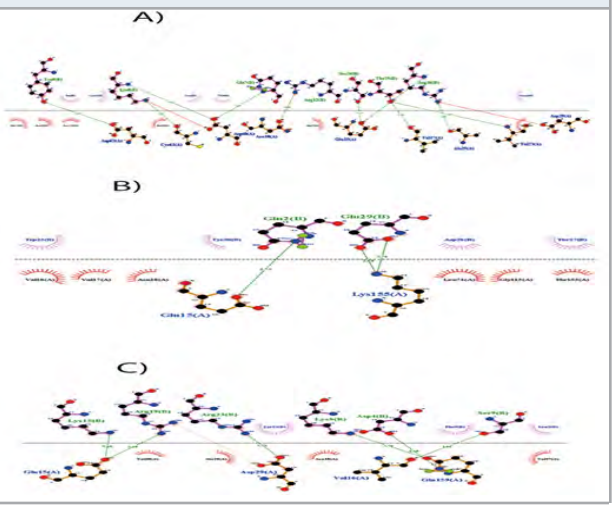


Figure 6: Dimplot results of FimA docked with Antimicrobial peptides.chainA -FimA , chainB-AMPs a)fimA-HBD3 b) fimA- Hevein c) fimA-LL-37



(A):LYS15(B),ARG19(B),of LL-37. it can be seen in the Dimplot(Figure 6).

CONCLUSION

The best docked complex is dermcidin and tobramycin with an affinity of - 5.2 and FimA with tetracyclin with an affinity of -5.9. This study has few limitations. One limitation of this study is that molecular level analysis of the docked complexes cannot be done since high resolution techniques such as X-ray crystallography should be done to verify the in-vivo complex formation of AMPs and antibiotics. The other limitation is, that electrophoresis technique is required to verify the increase in the molecular weight of docked complexes.

In addition, further experimental investigation is required to verify the binding of already docked complexes with matrix polysaccharides and proteins.

REFERENCES

Busch, A., Phan, G. and Waksman, G. (2015) 'Molecular mechanism of bacterial type 1 and P pili assembly', Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 373(2036). doi: 10.1098/rsta.2013.0153.

Chileveru, H. R. et al. (2015) Visualizing attack of escherichia coli by the antimicrobial peptide human defensin 5, Biochemistry. doi: 10.1021/bi501483q.

Dhople, V., Krukemeyer, A. and Ramamoorthy, A. (2006) 'The human beta-defensin-3, an antibacterial peptide with multiple biological functions', Biochimica et Biophysica Acta - Biomembranes. doi: 10.1016/j.bbamem.2006.07.007.

Giltner, C. L., Nguyen, Y. and Burrows, L. L. (2012) 'Type IV Pilin Proteins: Versatile Molecular Modules', Microbiology and Molecular Biology Reviews, 76(4), pp. 740–772. doi: 10.1128/mmbr.00035-12.

Govindarajan, D. K. et al. (2020) 'Adherence patterns of Escherichia coli in the intestine and its role in pathogenesis', Medicine in Microecology. doi: 10.1016/j.medmic.2020.100025.

Guillermo Garcia-Manero Shao-Qing Kuang, Susan O'Brien, Deborah Thomas, and Hagop Kantarjian, H. Y. (2005) 'NIH Public Access', Bone, 23(1), pp. 1–7. doi: 10.1016/j.tim.2010.03.002.A.

Kandaswamy, K. et al. (2013) 'Focal targeting by human β-defensin 2 disrupts localized virulence factor assembly sites in Enterococcus faecalis', Proceedings of the National Academy of Sciences of the United States of America, 110(50), pp. 20230–20235. doi: 10.1073/pnas.1319066110.

La Rosa, S. L. et al. (2016) 'Enterococcus faecalis ebp pili are important for cell-cell aggregation and intraspecies gene transfer', Microbiology (United Kingdom), 162(5), pp. 798–802. doi: 10.1099/mic.0.000276.

Lemkul, J. (2019) 'From Proteins to Perturbed Hamiltonians: A Suite of Tutorials for the GROMACS-2018 Molecular Simulation Package [Article v1.0]', Living Journal of Computational Molecular Science. doi: 10.33011/livecoms.1.1.5068.

Madhavi Sastry, G. et al. (2013) 'Protein and ligand preparation: Parameters, protocols, and influence on virtual screening enrichments', Journal of Computer-Aided Molecular Design. doi: 10.1007/s10822-013-9644-8.

Madrazo, M. et al. (2020) 'Predictive factors for Enterococcus faecalis in complicated community-acquired urinary tract infections in older patients', Geriatrics and Gerontology International, 20(3), pp. 183–186. doi: 10.1111/ggi.13856.

Muschiol, S. et al. (2019) 'Gram-Positive Type IV Pili and Competence', Protein Secretion in Bacteria, pp. 129–135. doi: 10.1128/microbiolspec.psib-0011-2018.

Overhage, J. et al. (2008) 'Human host defense peptide LL-37 prevents bacterial biofilm formation', Infection and Immunity. doi: 10.1128/IAI.00318-08.

Pamp, S. J. et al. (2008) 'Tolerance to the antimicrobial peptide colistin in Pseudomonas aeruginosa biofilms is linked to metabolically active cells, and depends on the pmr and mexAB-oprM genes', Molecular Microbiology. doi: 10.1111/j.1365-2958.2008.06152.x.

Prabhu, S. et al. (2013) 'Anionic Antimicrobial and Anticancer Peptides from Plants', Critical Reviews in Plant Sciences. doi: 10.1080/07352689.2013.773238.

Schittek, B. et al. (2001) 'Dermcidin: A novel human antibiotic peptide secreted by sweat glands', Nature Immunology. doi: 10.1038/ni732.

Schmidtchen, A. et al. (2002) 'Proteinases of common pathogenic bacteria degrade and inactivate the antibacterial peptide LL-37', Molecular Microbiology. doi: 10.1046/j.1365-2958.2002.03146.x.

Sillanpää, J. et al. (2010) 'Characterization of the ebpfm pilus-encoding operon of enterococcus faecium and its role in biofilm formation and virulence in a murine model of urinary tract infection', Virulence. doi: 10.4161/viru.1.4.11966.

Sivaramakrishnan, M. et al. (2019) 'Screening of curcumin analogues targeting Sortase A enzyme of Enterococcus faecalis: a molecular dynamics approach', Journal of Proteins and Proteomics. doi: 10.1007/s42485-019-00020-y.

Telford, J. L. et al. (2006) 'Pili in Gram-positive pathogens', Nature Reviews Microbiology, 4(7), pp. 509–519. doi: 10.1038/nrmicro1443.

Thappeta, K. R. V. et al. (2020) 'Combined Efficacy of an Antimicrobial Cationic Peptide Polymer with Conventional Antibiotics to Combat Multidrug-Resistant Pathogens', ACS infectious diseases, 6(5), pp. 1228–1237. doi: 10.1021/acsinfecdis.0c00016.

Tina, K. G., Bhadra, R. and Srinivasan, N. (2007) 'PIC: Protein Interactions Calculator', Nucleic Acids Research, 35(SUPPL.2), pp. 473–476. doi: 10.1093/nar/gkm423.

Tseng, J. T., Bryan, L. E. and Van den Elzen, H. M. (1972) 'Mechanisms and spectrum of streptomycin resistance in a natural population of Pseudomonas aeruginosa.', Antimicrobial agents

Biosc.Biotech.Res.Comm. Special Issue Vol 13 No 11 (2020) Pp-52-54

BBRC
Bioscience Biotechnology Research Communications

Malarial Parasite Identification Using Convolution Neural Network

S.Kavitha[1], S. Sathyavathi[2], R. Priyadharshini[3] and S.Varshini[4]
[1]*Assistant Professor, Department of Information Technology, Kumaraguru College of Technology, Coimbatore, India.*
[2]*Assistant Professor, Department of Information Technology, Kumaraguru College of Technology, Coimbatore, India.*
[3,4]*Student, Department of Information Technology, Kumaraguru College of Technology. Coimbatore, India.*

ABSTRACT
Malaria - a dreadful and deadly disease caused by a parasite belong to the plasmodium family that commonly infects a female Anopheles mosquito which bite on humans. With the symptoms, the disease can be diagnosed by trained lab technicians who will examine the microscopic blood smear images. Developing an automatic, accurate and efficient model for detecting this disease will reduce the requirement for the trained human resource and it will improve the diagnosis efficiency. Deep learning neural networks can be used to improve the efficiency and the accuracy of the diagnosis. In this paper, we propose a model using Convolutional Neural Network (CNN) for the examination of malaria from the microscopic human red blood smear images. This model will provide a rapid, accurate, low cost outcome. Our model differentiates the infected and uninfected cell images by training the convolutional neural networks. The algorithm involves the methods and architectures of computer vision, image processing operations and deep learning. The proposed CNN model can examine the malarial parasites from microscopic images with an accuracy of 68.38%, in 10000 checkpoint operations.

KEY WORDS: BLOOD SMEAR, CNN, DEEP LEARNING, MICROSCOPIC, PLASMODIUM PARASITES.

INTRODUCTION

Malaria is infectious and deadly disease. It is caused by the Plasmodium parasites. Malaria- mosquito-borne disease are spread by the bites of the infected female Anopheles mosquitoes. Out of many only 5 *Plasmodium* species cause malaria in human. They are *Plasmodium falciparum, Plasmodium vivax, Plasmodium malariae, Plasmodium ovale*, and *Plasmodium knowlesi*. If an infected female Anopheles mosquito bit a person, parasites in the mosquito emerge into human blood of the person and start destructing the oxygen-carrying human red blood cells (RBC). The symptoms are visibly noted in a few days or a week after the mosquito bites. Initial symptom is with the fever.

Without causing any symptoms the parasites can live in the human body over a year also. The delay in treatment for malaria can lead to multiple complications in the human body or sometimes it leads to death. Early detection of this dreadful disease can save life of the human. World Health Organisation (WHO) confirm that the diagnosis of the disease involves careful testing of the blood smear at 100X magnification. Lab technicians test the blood and count how many red blood cells contain parasites for every 5,000 cells. This process is complex.

ARTICLE INFORMATION

*Corresponding Author: sathyavathi.s.it@kct.ac.in
Received 4th Oct 2020 Accepted after revision 10th Dec 2020
Print ISSN: 0974-6455 Online ISSN: 2321-4007 CODEN: BBRCBA

Thomson Reuters ISI Web of Science Clarivate Analytics USA and Crossref Indexed Journal

NAAS Journal Score 2020 (4.31) SJIF: 2020 (7.728)

Online Contents Available at: http//www.bbrc.in/
Doi: http://dx.doi.org/10.21786/bbrc/13.11/12

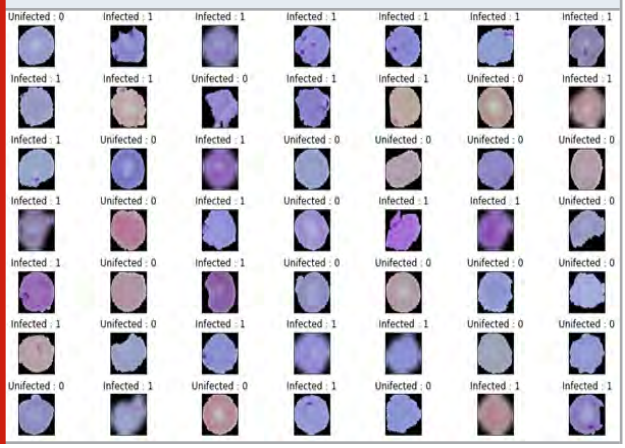
To improve the accuracy deep learning models can be used which will save the human life.

METHODOLOGY

Publicly available malaria dataset is used to conduct the experiments. The following are the steps involved: Importing Libraries, Loading the data, Data pre-processing, Data augmentation, Plotting images and its labels to understand how does an infected cell and uninfected cell looks like, Splitting data in Train, Evaluation and Test set, Creating a Convolution Neural Network function, Wrapping it with TensorFlow Estimator function, Training the data on Train data, Evaluating on evaluation data, Predicting on Test data, Plotting the predicted image and its respective True value and predicted value.

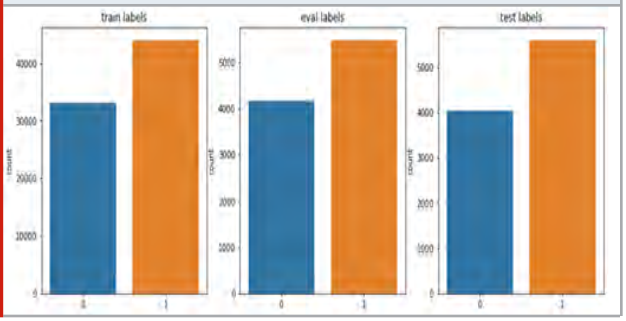
Dataset Collection: The dataset is taken from the publicly available official Kaggle Website. Total of 27,558 cell images are taken which comprises of both parasitized and uninfected cell images. Number of parasitized and uninfected cell images are equal. Reduction of images is done by carefully by annotating the images, where we remove the falsely and suspicious images, Number of false and suspicious images are about 647. Removing that from the data set which results in 26,161 images. Out of 26161 images, parasitized data stands 13,132 and the remaining are uninfected cell images.

Figure 1: Pre-processed Image Dataset



Data Pre-Processing: Data pre-processing is a basic step as the quality of image and the useful information that can be extracted from the image that directly affects the performance of the model to learn and predict. The behaviour and performances of the model purely depends on the data. After the image data acquisition process, most deep learning models perform various pre-processing methods.This would enhance the quality of the image by eliminating the noise, enhancing theillumination, improvising the colour variations inherent in the collection and staining phase of the image. So, data pre-processing is an important and essential process in any deep learning experiments. In this experiment, images are manually corrected. Images are resized and image patches are rescaled to map the features between 0 and 1 range as per the model requirement which leads to faster convergence. Data augmentation was applied to training data to improve the model performances. Pre-processing is done to improve the image quality and to reduce differences in the images that would avoid the complications in the forthcoming operations.

Figure 2: Dataset after splitting



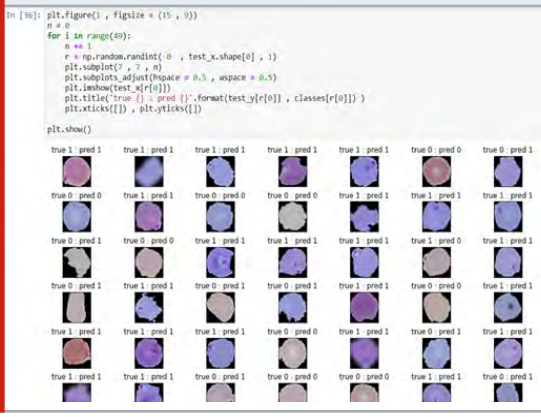
Data Splitting: Split the dataset into three sets in the ratio of 80(training):10(validation):10(test).We used sci-kit (open source learn machine learning library and imported "train_test_split" from sklearn_model_selection which splits arrays or matrices into random train and Test subsets. The Random state ensures that the splits generated are reproducible. The provided random state is used as a seed to the random number generator. This ensures that the random elements are generated in the same order. To ensure the train-test splits are always deterministic, random state which set a seed for the random generator is used. If the seed is not set properly, train-test split will be not same at all time.

Convolutional Neural Network: Inspired by the animal visual system, CNN is an artificial neural network. CNNs can extract features compared to conventional neural networks without losing a lot of spatial input correlations. Each layer consists of neurons that contain weights and prejudices that can be learned. After feeding data into the network and minimising the loss function at the top layer, the optimal model is acquired. A filter of value 50 with padding set to the same is used for convolution over the input volume in the convolutional layer's operation. In every max pooling layers size of 2 x 2 pool is used. The flatten layer with output neurons receive the input as the output from the final max pooling layer. Next is the two blocks of dense layer of output neurons. Drop out layer is used which discards 50% of the input neurons in a random manner.

The activation function Rectified Linear Unit (ReLU) is used in the convolution layers. SoftMax activation is used in the final stage of classification. Its output represents the vector which shows the probability distribution of the outcomes. The loss function sparse SoftMax is used to calculate the error between the actual value and predicted output value probabilities. Gradient Descent Optimizer is used to ensure the adaptive learning process. It optimizes the biases and weights of the network. Initially the weight

is assigned randomly, and the biases are initialised to 0. The activation function are applied for the batch size of 100 samples.

Figure 3: Execution result

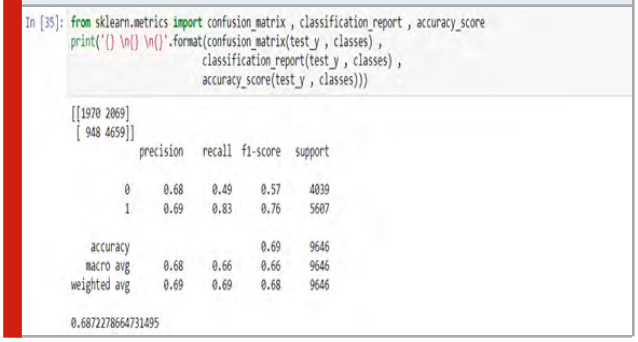


Confusion Matrix: The number of correct and incorrect predictions made by a classifier is evaluated. The summary of the predictions either correct or incorrect is given in the table named as confusion matrix. The performance of a classification model is calculated using the metrics accuracy, precision, recall, and F1-score.

CONCLUSION

In this work, we directed a progression of tests dependent on deep learning models to make efficient malarial parasite classification from classified human red blood cell smear images. We show that data augmentation methodologies used on the training set shows definite results. Our model achieves an accuracy of 68.38%. In future works, the different deep learning architecture can be used and analysed to understand to improve the accuracy.

Figure 4: Confusion Matrix



REFERENCES

Confusion Matrix Guide : https://www.educative.io/edpresso/how-to-create-a-confusion-matrix-in-python-using-scikit-learn

Mahdieh Poostchi .Image analysis and machine learning for detecting malaria by Kamolrat Silamut , Richard J.Maude , Stefan Jaeger , George Thoma , Volume 194, (2018).

Muhammad Umer , Saima Sadiq Muhammad Ahmad, Saleem Ullah, Gyu Sang Choi, Arif Mehmood (2020) PP(99):1-1

Muhammad Umer. A Novel Stacked CNN for Malarial Parasite Detection by Pages 36-55.

tf.compat.v1.train.GradientDescentOptimizer Guide : https://www.tensorflow.org/api_docs/python/tf/compat/v1/train/GradientDescentOptimizer

World Health Organization. Malaria: fact sheet. No. WHO-EM/MAC/035/E. World Health Organization. Regional Office for the Eastern Mediterranean (2014).

Biosc.Biotech.Res.Comm. Special Issue Vol 13 No 11 (2020) Pp-55-60

BBRC
Bioscience Biotechnology Research Communications

Towards Improving Skin Cancer Detection Using Transfer Learning

S. Sasikala[1]*, S. Arun Kumar[2], S.N. Shivappriya[3] and Priyadharshini T[4]
[1,3]Assosciate Professor, [2]Assistant Professor, [4]UG Scholar
Department of Electronics and Communication Engineering,
Kumaraguru College of Technology, Coimbatore-49, India

ABSTRACT

In present time, skin cancer is the deadliest disease among humans. In US, two persons die every hour owing to skin cancer. Skin cancer is developed on the body when exposed to sunlight and is the abnormal growth of the skin cell. The patient's life can be saved through earlier and faster detection of skin cancer. The formal method of skin cancer detection is Biopsy, it is done by removing the skin cells and testing the samples in a clinical lab. Biopsy method is invasive and time-consuming. With the newer technologies, early detection of skin cancer at the initial stage is possible. Image processing techniques are instrumental in the health care industry to detect abnormalities in the human body. In this work, Convolutional Neural Network (CNN) algorithm with four different transfer learning techniques are used to classify the images of the skin with dermoscopic analysis which enables fast detection. A CNN model is trained using a dataset of 3700 clinical images and its performance is tested over 660 images which represent the identification of deadliest skin cancer. A considerable improvement in accuracy of skin cancer detection using deep learning architecture ResNet34 provides a reliable approach for early detection and treatment.

KEY WORDS: CONVOLUTIONAL NEURAL NETWORK, DEEP LEARNING, DETECTION, TRANSFER LEARNING, SKIN CANCER.

INTRODUCTION

Skin cancer is the uncontrollable growth of damaged cells in the outer most layer of the skin. This is because of damage in DNA sequence due to the environmental factors like cigarette smoke and exposure to Ultra Violet (UV) light. DNA damage triggers mutation which leads to rapid multiplication of skin cells that forms malignant tumors [Miller .et. al. (1994)].

Skin cancer is classified into Melanoma, Basal Cell Carcinoma (BCC) and Squamous Cell Carcinoma (SCC). Melanoma is the most dangerous type of cancer which leads to death that usually appears on the moles and the areas on the skin which is exposed to sunlight as well as not exposed to sunlight. The affected part of the skin contains melanocytes that spread to other parts of the body. BCC is the most laggard growing and never be large in size. It appears on the skin exposed areas such as hand, face, leg, ears and scalp. It usually matures as an ulcer and does not improve. The early detection of this can be curable. Some are hostile and cannot be treated because it spreads to the deeper cells of the tissue. SQC appears on the sun exposed part and on the incurable inflammation part of the body and occurs in the person who has low immune power. It is large, appears in incurable scars and in lips. The early detection is possible.

ARTICLE INFORMATION

*Corresponding Author: sasikala.s.ece@kct.a.in
Received 05th Oct 2020 Accepted after revision 10th Dec 2020
Print ISSN: 0974-6455 Online ISSN: 2321-4007 CODEN: BBRCBA

Thomson Reuters ISI Web of Science Clarivate Analytics USA and Crossref Indexed Journal

NAAS Journal Score 2020 (4.31) SJIF: 2020 (7.728)

Online Contents Available at: http//www.bbrc.in/
Doi: http://dx.doi.org/10.21786/bbrc/13.11/13

Benign is a non-cancerous which does not spread to any other parts of the body. It is caused due to exposure of sunlight, inflammation of skin, infections, and genetics.

Melanoma mostly occurs in the skin rarely in the mouth and intestines with the abnormal cells that contain melanocytes which control the pigment in our skin. For women, melanoma mostly occurs on the legs and for men on the back. They usually develop from the mole with abnormal changes as an increase in size, changes in the color, causes itches or skin breakdown. It can occur in the areas between fingernails, palms, toenails and eyes [Miller .et. al. (1994)]. Benign usually appears on the skin which is highly exposed to sunlight such as face, shoulders, neck, hand and leg. This appears as lump and looks like patches which continues after a week and develops over a month or a year.

Skin cancer is the common type of cancer in worldwide and especially in US. By the age of 70, skin cancer will be developed by 1 in 5 Americans. In every hour, more than 2 people die because of skin cancer. Risk for melanoma will be doubled while exposure of sunburns is more than 5 in number. Early detection helps to survive for 5 years and the survival rate is 99 percent. At least 40% of cases have skin cancer when globally accounting for common cancer. Non-melanoma skin cancer occurs 2 to 3 million people per year. Globally in 2012, 232,000 people were in skin cancer, and 55,000 people died. According to the survey of last 20 to 40 years, Australia (white people), New Zealand and South Africa People have the highest rate of Skin cancer in the world [Apalla, Z.et. al. (2017)].

The early detection involves Biopsy method, in that the damaged skin is removed and tested in laboratories which take longer duration for the detection of skin cancer and it is more painful method. Computer Aided Diagnosis (CAD) is used to avoid the longer duration consumption and it is a non-invasive method. Many works in literature [Sasikala, S. et al. (2018) & (2020)] have focused on machine learning, feature transformation, optimisation, and deep learning for improving the cancer diagnosis. Henceforth, the proposed work aims to develop a CAD system for early identification of skin cancer using deep learning.

The significant contributions of the proposed work are:

- To design a cost-effective CAD system for the early identification of skin cancer.
- To construct a deep learning model that can detect and categorize given image into either a benign type or malignant type.
- To reduce training time by using pre-trained model with high accuracy rate.

Related Works: Automatic detection of skin cancer involved pre-processing and the post-processing techniques for the classification of the image with high accuracy. The Pre-processing increases the performance by noise removal. Post-processing enhanced the image quality and the boundary of the cancer cell was enhanced. The problem is that the dataset was too small. There was no uniqueness in the image processing technique and the variations between the dermoscopy and the digital image were large. By increasing the number of images in the data set, high accuracy over training and testing data set could be obtained [Lau, H.T. and Al-Jumaily, A., (2009)].

Skin cancer is the most common disease in human and its incidence is increasing dramatically. The newer technology-based detecting skin cancer is recommended for accurate identification but the percentage of detection by computer is comparatively high with manual detection. Here Digital Dermoscopy is widely considered as one of the most effective means to classify the skin cancers. Segmentation of images is done using K-means algorithm. It includes various stages like skin image, enhancement, lesion segmentation, feature extraction and finally classifying it to normal and abnormal. Convolution Neural Network (CNN) for detection of images, which is much cost-efficient in comparison with digital dermoscopy gives accurate output with appropriate detection of the cancer. The model proposed the automatic method of detecting skin cancer from the photographed image which was captured from the affected area of the skin. In this, Support Vector Machine (SVM) algorithm is used to classify the image which was either melanoma or benign. A total of 100 image for melanoma and 100 for benign was used in the study. The eminent drawback of this method is that the data set did not contain the dark-skinned images. To overcome these difficulties, more images are used to train the data and to test the trained data with all types of images of the skin [Esteva, A et. al (2017)].

Three methods were used for the segmentation of Otsu's method, gradient vector flow and color-based segmentation using K-mean clustering. Otsu's method did not require any variation in the parameter for the different skin lesions. The gradient vector flow had the active contour to boundary concavities although with the presence of the noises. The drawback of it was the execution speed to converge to the object. This method required the changing parameter for the different skin lesions. In color-based classification, it had the possibility of reducing the computational cost calculation for every pixel in the image. To overcome these drawbacks, the Convolutional Neural Network (CNN) algorithm with the back-propagation model is used for the fast execution to converge the object image for training [Rubegni P et al. (2012)].

An artificial neural network algorithm was used for the classification of the image and feature extraction for the thresholding. This method involved a binary classifier for classifying benign and malignant tumor. By the result of the artificial neural network, the accuracy was low. The different types of classification methods and image

processing techniques could be used for high accuracy and the accurate detection of the cancer [Hosny, K.M., Kassem, M.A. and Foaud, M.M. (2018)].

Melanoma and benign lesions have high similarity, due to this it takes a long time to identify and classify. The automatic classification of skin lesions helps to reduce time, efforts and one of the best ways to give an accurate identification of lesions. The use of transfer learning and pre-trained neural network has been experimented. In this method, binary classifier model was used. This proposed method provides accuracy around 96 percentage. The weights are fine-tuned and the dataset undergoes various rotation angles to overcome problems. From this paper, we included a feature for greater accuracy for the better identification of lesions [Mahmoud, M.K.A., Al-Jumaily, A. and Takruri, M (2011)].

Rule based approach, back propagation and neural network was used to select the features and to classify the lesions. The lesions were either melanoma or benign. As in this, the number of correct classifications increased. The neural network handled the complex relations among the identification of lesions. This model had a drawback of slow convergence rate and the trapping of the local minima. To overcome this, CNN could be used to increase the convergence rate and reduce the complexity of classifying the images to increase the trapping of the local minima [Mendes, D.B. and da Silva, N.C. (2018)]. ResNet-134 architecture was used that was trained over 3797 images and later 956 images were tested with the network and achieved an accuracy of about 78%. This technique took a long time to train the data set of approximately 35 hours. The images were not trained properly and led to the wrong prediction of the cancer lesions and made a problem for humans. By using ResNet 34, the time consumption to train the data set and the accuracy could be reduced [Jain, S. and Pise, N. (2015)].

SVM algorithm and snake active contour were used for image segmentation. To reduce the complexity in SVM, the snake parameters were used to predict the initial curve. The segmentation and the classification of the images are not accurate. To overcome these difficulties, CNN with ResNet 34 architecture could be used to detect the image accurately without any discrepancy [Aswin, R.B., Jaleel, J.A. and Salim, S. (2014)]. Computer-aided approach was used to detect skin cancer. The steps involved in the detection of cancer are image pre-processing, segmentation, feature extraction, and classification. Image resizing affected the quality of classification. By using this method, segmentation process involved the drawback was that the large stacks of data set could not be used for the classification of the image. To overcome these drawbacks, the CNN algorithm is used to classify large data set for accurate results [Hosny, K.M., Kassem, M.A. and Foaud, M.M.(2019)].

Transfer learning was applied with AlexNet by replacing last layer to classify the different types of lesions as melanoma and benign by softmax classification. CNN algorithm was used in this method to classify the images as melanoma or benign. The performance accuracy of the proposed model was high. The back-propagation method was used to fine-tune the weights to classify the images accurately. From the literature, it is observed that using CNN with the ResNet 34 architecture gives high accuracy with the low error rate with the probability of the classifier. Therefore, the proposed method for classifying benign and malignant tumors using CNN with ResNet architecture.

MATERIAL AND METHODS

In the present time, machine learning and deep learning approaches are used in healthcare for improved diagnosis. In this work, Convolutional Neural Network (CNN) algorithm with four different transfer learning approaches viz, AlexNet, VGG16, ResNet50, ResNet34 are used to classify the images of the skin with dermoscopic analysis which enables fast detection. The workflow with RESNET 34 transfer learning is proposed in figure 1.

Figure 1: Architecture for Skin Cancer detection

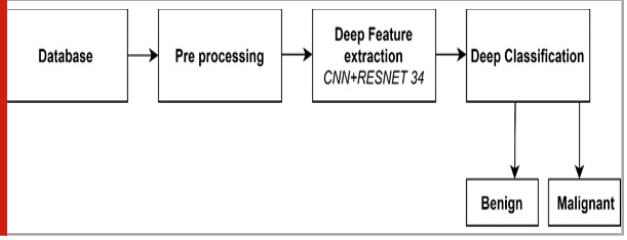
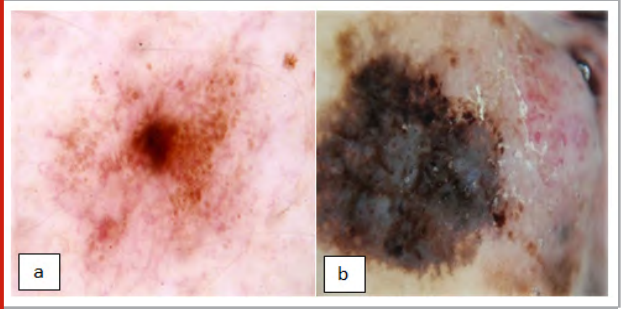


Table I represents the number of layers and the parameters used in the four transfer learning approaches used in this work.

Table 1. Different types of CNN Architectures Used

Name of the Architecture	Number of Layers	Number of Parameters
Alexnet	8	61M
VGG 16	16	138 M
ResNet 34	34	21.282M
ResNet 50	50	23.521M

Figure 2: Benign tumour Malignant tumour

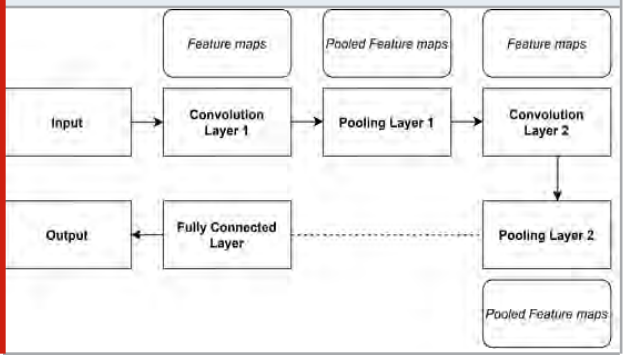


Data Base: The data sets are clinical images which are collected from Kaggle [kaggle.com]. The dataset has images of benign and malignant skin lesions in a balanced number. It consists of malignant and benign classes with 1800 images each. 70% of data is used for training and 30% for testing. A sample image in database is shown in figure 2.

Transfer learning is one of the machine learning techniques, used to develop a model for any recognition/ classification task. Also, it is the recommended method in deep learning where pre-trained models are used as the starting point of analysis. In transfer learning, a model trained on a single task is repurposed on another related task. It is an optimization that allows quick progress when modelling the new task.

Convolutional Neural Network (CNN): A Convolutional Neural Network (CNN or ConvNet) is a specialized type of artificial neural network (ANN) which is used in image processing, recognition and this is designed to process pixel data. CNN has some layers as the input layer, an output layer and hidden layers (multiple convolutional layers, pooling layers, fully connected layers and normalization layers) [Krizhevsky, A., Sutskever, I. and Hinton, G.E(2017)]. The figure 3 represents the highly performing architecture of CNN. Input Layer. The input layer or volume is an image and the dimensions of the image are [width x height x depth]. This dimension denotes the matrix of pixel values. For example, input is [32x32x3]. So width=32, height=32 and depth=3. Here the depth represents R, G, B channels. And the input volume should be divisible number of times by 2.

Figure 3: Convolutional Neural Network with its layers



Convolutional Layer: The main objective of this layer is feature extraction from the input layer. A small part of the image is joined to the convolutional layer to reduce the expense. For this purpose, dot products are applied between a filter and a receptive field on all the dimensions. After this dot product, the output volume with a single integer is obtained. This is known as a feature map. This process is done for the entire input image. The input for the next layer will be the output of the present layer. For feature detection, filter, kernel or feature detector which is a small matrix is used. The size of the typical filter on the first layer of CNN is [5x5x3]. After computing the dot product and by sliding the filter over the image, a convolved feature, activation map or feature map is formed (output volume). The number of filters is known as depth. The size of the filter and the receptive field, which is the local region of the input volume are the same. The set of neurons that are all pointing to the receptive field is known as depth column or fiber. Stride is used for spatially producing smaller output volumes.

Pool Layer: The function of this layer reduces the computational complexity of the model and spatial dimensions of the given input data. Over fitting is also controlled by this layer. It does not depend on the depth slice of the input. Different functions are Max pool, Average pool or L2 norm-pool. Max pooling is the most important part of the input layer.

Table 2. Results obtained with four CNN models

MODEL/ Metrics	ResNet34	ResNet50	VGG16	AlexNet
Accuracy	88.4 %	76.9 %	69.2 %	65.3 %
Error Rate	17.6 %	30.7 %	26.9 %	26.9 %
Sensitivity	0.875	0.875	0.875	1
Specificity	0.778	0.875	0.556	0.5
F1 score	0.736	0.736	0.608	0.64

Table 3. Epoch Vs accuracy and error rate table

EPOCH	TRAINING LOSS	VALID LOSS	ACCURACY (%)	ERROR RATE
0	0.8663	0.5069	79.78	0.202
10	0.4237	0.3699	84.49	0.1550
20	0.2985	0.3901	81.76	0.1823
30	0.2321	0.3265	86.77	0.1322
40	0.1562	0.3912	86.02	0.1398
50	0.1219	0.3808	87.84	0.1215
60	0.1120	0.3591	88.14	0.1185
70	0.084	0.3691	88.45	0.1155
80	0.0655	0.3957	88.90	0.1109
90	0.0486	0.3945	88.75	0.1124
99	0.0661	0.3858	88.6	0.1139

Fully Connected Layer: The main function of this layer is to connect each neuron of one layer to each neuron in another layer. This layer uses the softmax activation function. This function is used for classifying the generated features of the input images into various classes. This classification is based on the training dataset.

Resnet: ResNet is the short form for a residual network which is an ANN kind. ResNet could train extraordinary deep neural networks with more than 500 layers and still achieves powerful performance. According to the number of layers in the network, the number of parameters is selected. At its core, it uses batch normalization. The input layer is adjusted by batch normalization for increasing the performance of the network. The problem of covariate shift is reduced. The identity connection helps to preserve the network from vanishing gradient problem. This connection is used in ResNet. Bottle design residual network is used by Deep Residual Network to increase the network performance [He, K., Zhang, X., Ren, S. and Sun, J. (2016)]. Res-Net converges faster than plain counterpart. It also reduces degradation problem. ResNet-34 architecture offers bigger batch size which reduces training time. Further, ResNet-34 has a better validation error of 5.6% when compared to VGG. This is due to different depths of models. Hence Resnet is used in the proposed work.

RESLTS AND DISCUSSION

The data sets are processed with pre-trained CNN ResNet 34 networks to classify the type of skin cancer that is either benign or malignant. The final testing stage is done by choosing random skin lesion image. It is then tested for accuracy and error rate. Skin cancer detection is done with python for Transfer learning. The initial results obtained with all four approaches are tabulated in the Table 2 below. Among the four models RESNET 34 shows better performance than the other three models.

ResNet 34 alters only the first and last layer that is classification layers for more efficiency when compared with already proposed methods. By using a pre-trained RESNET model and GPU, the training time had been reduced to 1 hour. The detection of Skin cancer involves binary classifier as malignant and benign. Skin cancer detection is done with a different number of Epochs and learning rates. The lower number of epoch and a higher learning rate results in more error probability and less accuracy. The number of epochs given and corresponding output are listed in the Table 3 for training ResNet using higher Learning Rate (LR). The loss vs accuracy obtained by initially testing with 100 images of data sets with 70 for train and validation respectively is studied. This resulted in high validation loss and less accuracy (62%). To improve the system accuracy the training process used several data sets. Further training with 4360 images (Train - 3700, Valid - 660) of data set gives improved accuracy of about 88.6% in 100 epochs.

For the accurate detection of skin cancer, the learning rate is given as lower and higher as 1e-06 and 1e-03 with a loss of 0.06 and 0.145 respectively. With the maximum learning rate of 1e-03, the following accuracy and error rate for 10 epochs are identified. This gives the maximum accuracy of 90.12% and the minimum error rate of 9.8% after 10 epochs. The table below shows the change in accuracy and error rate for every epoch. The results are tabulated in Table IV.

Table 4. Epoch Vs accuracy and error rate table for optimal learning rate

EPOCH	TRAINING LOSS	VALID LOSS	ACCURACY (%)	ERROR RATE
0	0.8663	0.3291	89.51	0.1048
1	0.4237	0.3354	89.96	0.1003
2	0.2985	0.3285	89.81	0.1018
3	0.2321	0.3951	89.66	0.1033
4	0.1562	0.3852	89.81	0.1018
5	0.1219	0.4383	89.05	0.1094
6	0.1120	0.3675	89.05	0.1094
7	0.084	0.4057	88.75	0.1124
8	0.0655	0.4074	88.75	0.1124
9	0.0486	0.4150	88.44	0.1155
10	0.0661	0.3712	90.12	0.0987

CONCLUSION

Detection of skin cancer lesions as malignant (melanoma) or benign is performed using the CNN RESNET34. The performance of this system is studied using the accuracy and error rate with respect to the variations in number of epochs and learning rate. The accuracy increases with decrease in learning rate. The Maximum accuracy of 90.12% is achieved when LR is decreased to 1x10-6 after 10 epochs. In this work, only the detection of skin cancer is considered. In future, it could be extended for the diagnosis of various types of skin cancer such as melanocytic nevi, BCC, SCC through multi class classifier and collecting appropriate data sets. The number of layers in the CNN could also be increased for further improvement in the performance. Strengthening the dataset by increasing the images in each class will also lead better diagnostic results.

REFERENCES

Apalla, Z., Nashan, D., Weller, R.B. and Castellsagué, X., (2017). Skin cancer: epidemiology, disease burden, pathophysiology, diagnosis, and therapeutic approaches. Dermatology and therapy, 7(1), pp.5-19..

Aswin, R.B., Jaleel, J.A. and Salim, S., (2014). Hybrid genetic algorithm–Artificial neural network classifier for skin cancer detection. In 2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT) (pp. 1304-1309). IEEE..

Miller, D.L. and Weinstock, M.A., (1994). Nonmelanoma skin cancer in the United States: incidence. Journal of the American Academy of Dermatology, 30(5), pp.774-778.

Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M. and Thrun, S., (2017). Dermatologist-level classification of skin cancer with deep neural networks.

nature, 542(7639), pp.115-118.

He, K., Zhang, X., Ren, S. and Sun, J., (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

Hosny, K.M., Kassem, M.A. and Foaud, M.M., (2018). Skin cancer classification using deep learning and transfer learning. In 2018 9th Cairo International Biomedical Engineering Conference (CIBEC) (pp. 90-93). IEEE.

Hosny, K.M., Kassem, M.A. and Foaud, M.M., (2019). Classification of skin lesions using transfer learning and augmentation with Alex-net. PloS one, 14(5), p.e0217293.

https://www.kaggle.com/fanconic/skin-cancer-malignant-vs-benign

Jain, S. and Pise, N., (2015). Computer aided melanoma skin cancer detection using image processing. Procedia Computer Science, 48, pp.735-740..

Krizhevsky, A., Sutskever, I. and Hinton, G.E., (2017). Imagenet classification with deep convolutional neural networks. Communications of the ACM, 60(6), pp.84-90.

Lau, H.T. and Al-Jumaily, A., (2009), December. Automatically early detection of skin cancer: Study based on neural netwok classification. In 2009 International Conference of Soft Computing and Pattern Recognition (pp. 375-380). IEEE.

Mustafa, S., & Kimura, A. (2018, January). An SVM-based diagnosis of melanoma using only useful image features. In 2018 International Workshop on Advanced Image Technology (IWAIT) (pp. 1-4). IEEE.

Mahmoud, M.K.A., Al-Jumaily, A. and Takruri, M., (2011). The automatic identification of melanoma by wavelet and curvelet analysis: study based on neural network classification. In 2011 11th International Conference on Hybrid Intelligent Systems (HIS) (pp. 680-685). IEEE..

Mendes, D.B. and da Silva, N.C., (2018). Skin lesions classification using convolutional neural networks in clinical images. arXiv preprint arXiv:1812.02316..

Rubegni P, Cevenini G, Burroni M, Perotti R, Dell'Eva G, Sbano P, Miracco C, Luzi P, Tosi P, Barbini P, Andreassi L. (2002) Automated diagnosis of pigmented skin lesions. International Journal of Cancer. 2002 Oct 20;101(6): 576-80.

Sasikala, S., Bharathi, M. and Sowmiya, B.R., (2018). Lung Cancer Detection and Classification Using Deep CNN. International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN, pp.2278-3075.

Sasikala, S., Ezhilarasi, M. and Kumar, S.A., (2020). Detection of breast cancer using fusion of MLO and CC view features through a hybrid technique based on binary firefly algorithm and optimum-path forest classifier. In Applied Nature-Inspired Computing: Algorithms and Case Studies (pp. 23-40). Springer, Singapore.

Biosc.Biotech.Res.Comm. Special Issue Vol 13 No 11 (2020) Pp-61-65

BBRC
Bioscience Biotechnology Research Communications

Deep Learning-based Image Analysis Model for Diagnosing Thyroid Carcinoma in Fine Needle Aspiration Cytology (FNAC) Images

Gopinath Balasubramanian[1] and Santhi Ramalingam[2]

[1]*Associate Professor, Department of Electronics and Communication Engineering, Kumaraguru College of Technology, Coimbatore-641049, India*

[2]*Assistant Professor, Department of Biochemistry, PSG College of Arts & Science, Coimbatore-641014, India*

ABSTRACT

In this work, the diagnostic accuracy of an automated diagnosis system is evaluated using two pre-trained convolutional neural network models, namely AlexNet and VGG16. The diagnosis system is used to identify the cancerous and normal thyroid cells in Fine Needle Aspiration Cytology (FNAC) photographs. The proposed Alexnet and VGG16 models are implemented using deep learning based Transfer Learning (TL) to process multi-stained FNAC images. Initially, the image patches are derived from the cytology images based on the thyroid cell population. These patches are fed to the 8-deep layered AlexNet and 16-deep layered VGG16 as inputs and they are passed through multiple convolution layers, max pooling and dense layers. Through optimal implementation and testing of the models, the AlexNet model achieves a diagnostic accuracy of 92.5% whereas the VGG16 model results a diagnostic accuracy of 96.66% and sensitivity and specificity of 98.75% and 92.5% respectively.

KEY WORDS: BENIGN, CARCINOMA, DEEP LEARNING, MALIGNANT, THYROID.

INTRODUCTION

Thyroid is a gland in human body producing Thyroid Hormone (TH) that regulates metabolic processes in human body for normal growth and development (Mullur et al., 2014). In few cases, Around the thyroid organ, unwanted tissue section is formed which is known as thyroid nodules. Normally, this swelling nature of nodule is harmless as well as treated as benign type of nodules. However, around 8–15% of them are accounted as malignant and the malignant nodules should be removed by suitable surgical treatment. (Cooper et al., 2009). Hence, accurate discrimination of malignant nodule is an essential thing. High degree of attention must be given for thyroid malignant studies (Kwong et al., 2015). FNAC is a simple cost-effective and widely used method procedure for evaluating the head and neck masses with an accuracy of around 90%. (Wong et al., 2020). In this procedure, the area to be aspirated is cleaned properly and a sufficient small size of biopsy sample is derived using a syringe of 23 gauge needle setup from the thyroid nodule. The smears are prepared by the pathologist using these samples. All the smears are stained with suitable staining protocol in the pathology laboratory and placed under a microscope to study the characteristics of cell structures of thyroid nodule. Based on the experience of the pathologist and nature of appearance of the smear

ARTICLE INFORMATION

*Corresponding Author: gopinath.b.ece@kct.ac.in
Received 11th Oct 2020 Accepted after revision 12th Dec 2020
Print ISSN: 0974-6455 Online ISSN: 2321-4007 CODEN: BBRCBA

Thomson Reuters ISI Web of Science Clarivate Analytics USA and Crossref Indexed Journal

NAAS Journal Score 2020 (4.31) SJIF: 2020 (7.728)

Online Contents Available at: http//www.bbrc.in/
Doi: http://dx.doi.org/10.21786/bbrc/13.11/14

under microscope, an appropriate diagnostic report is generated. However, this routine manual screening of cytological slides is a tedious task and subjective in nature.

Various automated techniques are currently used by the researchers for diagnosing malignant thyroid nodule. The latest developments in deep learning algorithms are being used to solve the issues in the diagnosis process (Thomas et al., 2020). A combined neural networks and morphometric feature-based model was designed for the discrimination problem and a diagnosis accuracy of 97.8% was achieved (Karakitsos et al., 1999). A k-NN classifier and Bayesian models were used by Würflinger et al., (2004) to classify the cell nuclei in pathology images and a diagnostic accuracy of 86.1% using Bayesian model and 87.5% using k-NN model were achieved. Ahmed et al., (2016) has proposed an intelligent diagnostic system for separating papillary and other thyroid carcinoma levels with neural network classifier model and obtained a diagnostic accuracy of 90.32%.

In the previous work, Gopinath, & Gupta, (2010) designed a computer aided diagnosis tool to differentiate cancerous and non-cancerous thyroid FNAC cytology images using Gabor features and reported 93.33% of diagnostic accuracy. Subsequently, authors (Gopinath, & Shanthi, 2013; Gopinath, & Shanthi, 2015) tested SVM, ENN, k-NN and decision tree classifiers trained with wavelet decomposition based statistical features, for the same problem, from which the single SVM and ENN classifiers resulted an accuracy of 90% to separate cancerous and non-cancerous thyroid nodules. However, an improved accuracy of 96.66% was achieved by combining SVM, ENN and k-NN classifiers using linear combination rule and majority voting rule. Using ultrasound radiology images also, there was an automated screening system developed for detecting the malignancy in Thyroid nodules. For the detection of nodules, initially, a region-based detection network was developed to extract pyramidal features. In the next stage, a classification network was developed to classify the ultrasound thyroid images. The detection and diagnostic accuracy of the deep-learning-based CAD system were observed as 97.5% and 97.1%, respectively (Liu et al., 2019).

The skin lesions were automatically classified using AlexNet model and transfer learning (Hosny et al., 2019). The weights of the architecture were fine-tuned and a softmax layer was used on skin lesions. The proposed method produced a classification accuracy of 97.70%. Chen et al., (2020) examined 345 thyroid sections by applying the deep learning techniques and extracting the patch features and 96.1% of classification accuracy was reported. A two-phase algorithm was developed for automated extraction of red blood cells (RBCs) (Aliyu et al., 2020). Initially, cell regions were extracted and in the second step, AlexNet was used for diagnosing the abnormal conditions in the cells. Around, 9000 images were processed, for the study, using AlexNet transfer learning model. The study resulted a prediction accuracy of 95.92%, sensitivity and specificity of 77% and 98.82% respectively. A kernel-based classifier and optimization techniques were used with significant feature choice by Shankar et al., (2020) which enhanced the diagnosis process and they achieved a classification accuracy of 97.49%.

The current research work proposes the evaluation of the diagnostic accuracy of an automated diagnosis system using two pre-trained convolutional neural network models, namely AlexNet and VGG16. The FNAC images are processed by implementing the neural network models to diagnose the normal and cancerous thyroid cell regions. The proposed Alexnet and VGG16 models are implemented using deep learning based Transfer Learning (TL) to process multi-stained FNAC images.

MATERIAL AND METHODS

Acquiring Thyroid FNAC images: Fine needle aspiration cytology is a partially invasive procedure for determining cancerous and normal state of the thyroid nodules under study. Thyroid biopsy samples of thyroid nodule are taken by a fine-needle and transferred to the pathology laboratory. Then, the glass slides are prepared and analyzed under a microscope by a pathologist. After the examination of slides, the pathologist generates a diagnosis summary of report. The diagnostic report consists of four classes, namely, malignant, benign, inadequate sample and suspicious state. In general, the FNAC technique produces an accuracy of around 90-95% whereas the falsely identified positive cases and falsely identified negative results are varying in the range of 0-5% (Sinna, & Ezzat, 2012).

From the report of FNAC, the Diagnostic Accuracy (DA), sensitivity and specificity parameters are obtained as given in Eqs. (1-3) respectively. The parameter sensitivity can calculate the count of truly positive (TP) that correctly determines a positive case as positive. Similarly, the specificity calculates the count of truly negative (TN) that correctly identifies a negative case as negative. The DA is a combined measure of true positive and true negative.

$$Sensitivity = \frac{TP}{TP+FN} \quad (1)$$

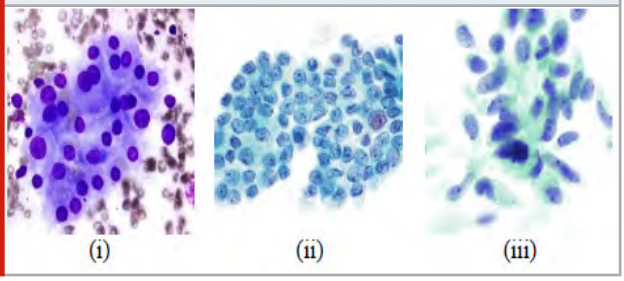
$$Specificity = \frac{TN}{TN+FP} \quad (2)$$

$$DA = \frac{(TP+TN)}{(FP+FN+TP+TN)} \quad (3)$$

Now, the FNAC cytology images are acquired focusing on the slide's portions of the biopsy sample with a help of a digital camera attached with the microscope. Then, the conventional manual screening method can be automated by developing automated diagnosis system. This system, whenever required, can be utilized as a tool for getting a second opinion to support the decision made by the pathologist. For testing the developed system, the thyroid FNAC cytological images are obtained from image atlas

database released by PapSociety which are evaluated and ratified by a panel of expert committee. The typical normal and cancerous FNAC images are shown in Figure 1. Figure 1(i) shows a benign image. Figures 1(ii) and Figure 1(iii) show papillary image and medullary image respectively.

Figure 1: (i) Benign, (ii) Papillary and (iii) Medullary Carcinoma FNAC images



Pre-trained CNN models: Krizhevsky et al., (2012) developed the AlexNet and it uses ImageNet which was presented by Deng et al., 2009. This research work uses 80 training images in which 40 images are belonging to benign and remaining 40 images are belonging to malignant group. On the other hand, 30 testing images are used in which 10 images are belonging to benign group and remaining 20 images are belonging to malignant group. All the images are manually segmented into 4 fragment patches which are having high concentrated cell population. Now, the total count of patch portions in training group is 320 images. Similarly, due to this manual segmentation, the total count of patches in testing group is 120 images as given in Table 1. These image segments are given as input to AlexNet and VGG16 pre-trained models.

Table 1. FNAC Image Set and its Patches with Training Set and Testing Set

	Total Images	*Image Patches*	*Benign*	*Image Patches*	*Malignant*	*Image Patches*
Training Set	80	320	40	160	40	160
Testing Set	30	120	10	40	20	80

The architecture of AlexNet is shown in Figure 2. The first layer of the AlexNet filters the input image. Then, the next convolutional layer receives the input from the filters of previous layer which is connected to the pooling layer. It keeps the significant features and reduces the large number of features into a smaller number of features. In Alex-Net, each layer has more filters. Like most convolutional neural networks, it has a combination of convolution layer and a pooling layer. It uses ReLU activation function which is more biological inspired. The third layer receives its input from the second convolutional layer. It has multiple kernels, and each kernel has a 3x3 size. Also, in fourth convolutional layer, 384 kernels are there. In this sequence, 256 kernels are there in the fifth convolutional layer with a same 3x3 kernel size. The convolutional layer outputs from third, fourth and fifth stages are given to two fully connected layers as inputs (Hosny et al., 2019).

Figure 2: Layout of pre-trained AlexNet Architecture

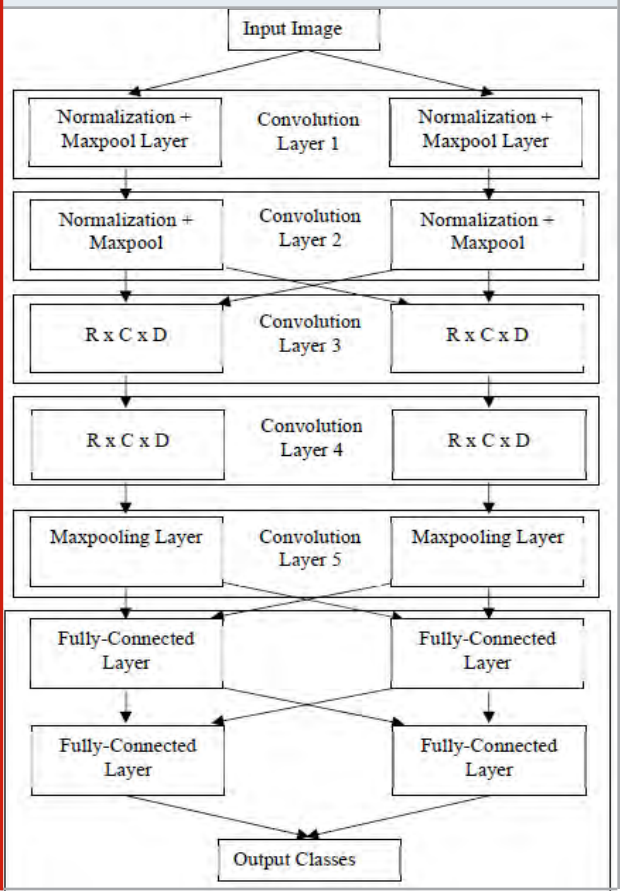
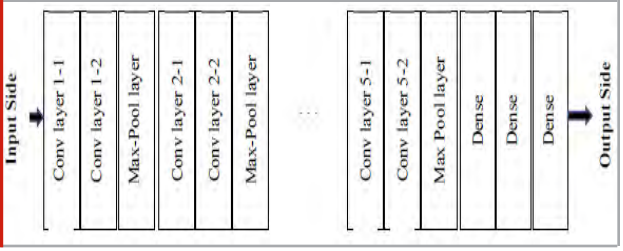


Figure 3 shows the complete layout of the VGG16 architecture. It has five blocks as a combination of two convolution and a max-pooling layers (Zhao et al., 2018; Belaid et al., 2020; Hameed et al., 2020). The first two blocks have two convolution layers and the 3rd, 4th and 5th blocks have three convolution layers. All the blocks are ended with a max-pooling layer. Each layer of VGG16 model uses the activation function 'ReLU'. Finally, the dense layers are being utilised for transferring neuron elements in the networking structure from input side to output side.

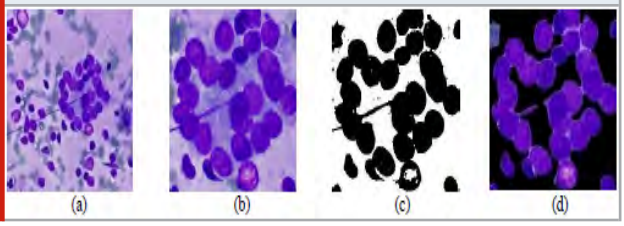
Figure 3: Layout of of pre-trained VGG16 architecture



RESULTS AND DISCUSSION

The proposed automated diagnosis system utilises the pre-trained models AlexNet and VGG16 with transfer learning. The input FNAC images are converted into patch segments. These patch images are pre-processed before the application of pre-trained models using Thresholding operation followed by segmentation of foreground cell regions. The thresholding operation is implemented using Otsu's algorithm on the input patches for the removal of background stain information and mathematical morphology segmentation procedure is used to further remove the small unwanted noise pixels present in the inner and outer sides of the thyroid cell portions.

Figure 4: (a) Input slide image of Thyroid FNAC, (b) Manually cropped patch portion, (c) Output of Thresholding operation and Mathematical morphology, (d) Superimposed image used to perform the classification



In mathematical morphology, the optimised size of disk-shaped structural element has been used to capture the foreground cell regions. The results are presented in Figure 4. The segmented thyroid FNAC image patches are now fed as the inputs to the proposed AlexNet and VGG16 pretrained models. These models are used to perform feature extraction and classification on the given input image patches and the result from these models are the required diagnosis results such as benign or malignant classes. The outcome of the pretrained architectures is evaluated by sensitivity, specificity and diagnostic accuracy.

Table 2. Performance analysis of AlexNet and VGG16

Parameters	AlexNet	VGG16
True Positive	75	79
True Negative	36	37
False Positive	5	1
False Negative	4	3
Diagnostic Accuracy	92.5%	96.66%
Sensitivity	93.75%	98.75%
Specificity	90%	92.5%

The results truly positive cases and truly negative cases are tabulated in Table 2 through which the other performance measures namely, sensitivity, specificity and diagnostic accuracy have been calculated. From Table 2, it is observed that both the models have performed well to diagnose the malignant thyroid cells in FNAC images. However, the VGG16 model has produced a higher sensitivity of 98.75%. This indicates that only one image patch is misdiagnosed in malignant group whereas five image patches are misdiagnosed by the AlexNet model. However, the VGG16 model has produced a specificity of 92.5% only. By analysing the overall performance of the two pre-trained models, the VGG16 is outperforming well in the diagnosis of thyroid FNAC pathology images with a higher diagnostic accuracy of 96.66%.

CONCLUSION

The performance of two pre-trained convolutional neural network models, namely AlexNet and VGG16 was evaluated to differentiate normal and cancerous thyroid cells in FNAC patch image sets. Both models were tested by implementing Transfer Learning concept in deep-learning based approach. In the initial step, the FNAC images were pre-processed and patch images were cropped based on the thyroid cell population. These patches were fed to the 8-deep layered AlexNet and 16-deep layered VGG16 as inputs and they were passed through multiple convolution layers, max pooling and dense layers. Through optimal implementation and testing of the models, the AlexNet model has achieved a diagnostic accuracy of 92.5% whereas the VGG16 model has reached a higher diagnostic accuracy of 96.66% along with 98.75% of sensitivity and 92.5% of specificity. From the results, it can be evaluated that the pre-trained AlexNet and VGG16 models can play a vital role in effective diagnosis of malignant carcinoma in cyto-pathology images.

REFERENCES

Ahmed J and Rahman MA (2016) Intelligent diagnostic system for Papillary Thyroid Carcinoma Journal of Applied Environmental and Biological Sciences, Vol 6 No 3 Pages 72-82

Aliyu HA Razak MAA Sudirman R and Ramli N (2020) A deep learning AlexNet model for classification of red blood cells in sickle cell anemia International Journal of Artificial Intelligence Vol 9 No 2 Pages 221-228

Belaid ON and Loudini M (2020) Classification of Brain Tumor by combination of pre-trained VGG16 CNN Journal of Information Technology Management Vol 12 No 2 Pages 13-25

Chen P Shi X Liang Y Li Y Yang L and Gader PD (2020) Interactive thyroid whole slide image diagnostic system using deep representation Computer Methods and Programs in Biomedicine Vol 195 Pages 105630

Cooper DS Doherty GM Haugen BR Kloos RT Lee SL Mandel SJ and Sherman SI (2009) Revised American Thyroid Association management guidelines for patients with thyroid nodules and differentiated thyroid cancer: the American Thyroid Association (ATA) guidelines

taskforce on thyroid nodules and differentiated thyroid cancer Thyroid Vol 19 No 11 Pages 1167-1214

Deng J Dong W Socher R Li LJ Li K and Fei-Fei L (2009) ImageNet: A large-scale hierarchical image database in Proceedings of IEEE Conference Computer Vision and Pattern Recognition Pages 248-255

Gopinath B and Gupta BR (2010) Majority voting based classification of thyroid carcinoma Procedia Computer Science Vol 2 Pages 265-271

Gopinath B and Shanthi N (2013) Computer-aided diagnosis system for classifying benign and malignant thyroid nodules in multi-stained FNAB cytological images Australasian Physical & Engineering Sciences in Medicine Vol 36 No 2 Pages 219-230

Gopinath B and Shanthi N (2015) Development of an automated medical diagnosis system for classifying thyroid tumor cells using multiple classifier fusion Technology in Cancer Research & Treatment Vol 14 No 5 Pages 653-662

Hameed Z Zahia S Garcia-Zapirain B Javier Aguirre J and María Vanegas A (2020) Breast cancer histopathology image classification using an ensemble of deep learning models Sensors Vol 20 No16 Pages 4373

Hosny KM Kassem MA and Foaud MM (2019) Classification of skin lesions using transfer learning and augmentation with Alex-net PloS one Vol 14 No 5 Pages e0217293

Karakitsos P Cochand-Priollet B Pouliakis A Guillausseau PJ and Ioakim-Liossi A (1999) Learning vector quantizer in the investigation of thyroid lesions Analytical and Quantitative Cytology and Histology Vol 21 No 3 Pages 201-208

Krizhevsky A Sutskever and Hinton G (2012) ImageNet Classification with Deep Convolutional Neural Networks In Proceedings of Neural Information Processing Systems (NIPS) Vol 1 Pages 1097-1105

Kwong N Medici M Angell TE Liu X Marqusee E Cibas ES and Alexander EK (2015) The influenceof patient age on thyroid nodule formation, multinodularity, and thyroid cancer risk The Journal of Clinical Endocrinology & Metabolism Vol 100 No 12 Pages 4434-4440

Liu T Guo Q Lian C Ren X Liang S Yu J and Shen D (2019) Automated detection and classification of thyroid nodules in ultrasound images using clinical-knowledge-guided convolutional neural networks Medical Image Analysis Vol 58 Pages 101555

Mullur R Liu YY and Brent GA (2014) Thyroid hormone regulation of metabolism Physiological Reviews Vol 94 No 2 Pages 355-382

Shankar K Lakshmanaprabu SK Gupta D Maseleno A and De Albuquerque VHC (2020) Optimal feature-based multi-kernel SVM approach for thyroid disease classification The Journal of Supercomputing Vol 76 No 2 Pages 1128-1143

Sinna EA and Ezzat N (2012) Diagnostic accuracy of fine needle aspiration cytology in thyroid lesions Journal of the Egyptian National Cancer Institute Vol 24 No 2 Pages 63-70

Thomas J Ledger GA and Mamillapalli CK (2020) Use of artificial intelligence and machine learning for estimating malignancy risk of thyroid nodules Current Opinion in Endocrinology, Diabetes and Obesity Vol 27 No 5 Pages 345-350

Wong CKH Liu X. and Lang BHH (2020) Cost-effectiveness of fine-needle aspiration cytology (FNAC) and watchful observation for incidental thyroid nodules Journal of Endocrinological Investigation Vol 43 Pages 1645-1654

Würflinger T Stockhausen J Meyer-Ebrecht D and Böcking A (2004) Robust automatic coregistration, segmentation, and classification of cell nuclei in multimodal cytopathological microscopic images. Computerized Medical Imaging and Graphics Vol 28 No 1-2 Pages 87-98

Zhao D Zhu D Lu J Luo Y and Zhang G (2018) Synthetic medical images using F&BGAN for improved lung nodules classification by multi-scale VGG16 Symmetry Vol 10 No 10 Pages 519

Biosc.Biotech.Res.Comm. Special Issue Vol 13 No 11 (2020) Pp-66-69

BBRC
Bioscience Biotechnology Research Communications

Evaluation of Wound Healing Capacity of Selected leaf Extracts using In vitro Scratch Assay with L929 Fibroblasts

Gowthama Prabu Udayakumar[1], Poorani Gurumallesh[1] and Baskar Ramakrishnan[1]*
[1]*Department of Biotechnology, Kumaraguru College of Technology, Coimbatore, Tamil Nadu 641049 India*

ABSTRACT

Wound healing is the process by which the skin repairs and maintains itself. Any delay in the healing might result in various skin pathologies like prolonged non-healing and chronic ulceration. Traditional medicines use plant-based products that play an important role in cutaneous wounds. L929 Fibroblast cell line forms the connective tissues, which hold importance by synthesizing extracellular matrix and collagen in the process of wound healing. In the present study, cold percolated ethanol leaf extracts of *Beta vulgaris* and *Psidium guajava* were compared for its in vitro wound healing activity through scratch wound assay performed on L929 cells. The rate of healing was examined at regular intervals and determined using ImageJ software. The skin cell re-epithelialization property was identified to be comparatively higher levels in the ethanol extract of Beta vulgaris when compared to that of *Psidium guajava*. This study aims to compare the in vitro wound healing activity of the selected ethanol extracts with that of the standard positive control, thereby extending its application for in vivo wound healing capacity useful in the cosmetic industries.

KEY WORDS: BETA VULGARIS; COLD PERCOLATION; FIBROBLAST; IMAGEJ; PSIDIUM GUAJAVA; WOUND SCRATCH ASSAY.

INTRODUCTION

Wound healing is a natural process of replacing the damaged cell or tissue with fresh ones. There are a chain of events occurring within a human body immediately post trauma and with the help of certain pathways, all the biomolecules would contribute to make the healing process a successful one (Sorg, Tilkorn, Hager, Hauser, & Mirastschijski, 2017). In adult human beings, the dynamic wound healing process is categorized into four different segments namely hemostasis phase, inflammatory phase, the proliferation phase and the remodeling phase. Generally, wound healing happens on its own and does not require any medical examinations. With respect to chronic health conditions like vascular disease and diabetes, the natural wound healing process is slightly impaired (Gosain & DiPietro, 2004). A majority of mechanisms through which healing occurs has been portrayed, still a lot of pathophysiological process are yet to be decoded.

Researchers identified certain factors to be extremely influential in hindering the natural process of wound healing and they were organized as local and systemic. Local factors have a straight impact on wound, however the systemic factors are related to overall health conditions. Single or multiple factors could be related to the healing process and its very complex to identify the trigger point that's impairing wound healing process (Guo & DiPietro, 2010). To overcome the hindrance

ARTICLE INFORMATION

*Corresponding Author: baskar.r.bt@kct.ac.in
Received 08th Oct 2020 Accepted after revision 14th Dec 2020
Print ISSN: 0974-6455 Online ISSN: 2321-4007 CODEN: BBRCBA

Thomson Reuters ISI Web of Science Clarivate Analytics USA and Crossref Indexed Journal

NAAS Journal Score 2020 (4.31) SJIF: 2020 (7.728)

Online Contents Available at: http//www.bbrc.in/
Doi: http://dx.doi.org/10.21786/bbrc/13.11/15

occurring in natural wound healing process, the aids play a very significant role. Extracted biomolecules or the synthetic drugs accelerated the wound healing process by effectively acting on the site on infection.

To regulate wound healing and to fasten it there are a lot of drug formulations being used by a lot of medical practitioners. The rise in human population, shortage in supply of drug molecules and multiple drug resistance are the major triggers pushing the researchers to find an effective alternative to sort out the problem. In multiple dimensions, there are approaches being done to accelerate the chronic wound healing. For an instance, a research group from China are exclusively working on RNA regulations that would facilitate the healing of diabetic chronic wounds (Z.-H. Liang, Pan, Lin, Qiu, & Zhang, 2020).

The synthetic medicines are offering a wide range of solution; still phytochemicals have its importance on its own. Curcumin contains a lot of bioactive molecules which are very significant in healing the wound. Also, herbal medicine is gaining attention worldwide from the perspective of both producer and user. To a range of patients, herbal medicines has been a supportive therapy / alternative therapy to treat a broad spectrum of ailments. With the emerging advancements purification technology and analytical chemistry, identifying and isolating the bioactive component has become more precise and reliable. Many phytochemicals were extracted multiple aspects and been formulated into various forms like a tablet / capsule / syrup and topical applications.

Understanding the biology of wound healing is absolutely complex as it involves a cascade of events and also many numbers of biomolecules are actively influential in the process (Thangapazham, Sharad, & Maheshwari, 2016). With progress in research towards phytochemicals in wound healing it was also found that it helps in skin regeneration along with treating the wound. Another important attribute to be taken into consideration is cost effectiveness. As natural resources are present in abundance, the production cost would be invariably lower and hence it could be reachable to all ranges of buyers with respect to commercialization. After a thorough literature review and analyzing previous researches made, this study mainly aims to compare and investigate the in vitro wound healing activity of the selected ethanol extracts of the leaves of *Beta vulgaris* and *Psidium guajava* on L929 Fibroblast cells using scratch wound assay. Based on the performance criteria and effectiveness, the scale up of the production could be maximized to meet up the demand.

MATERIAL AND METHODS

Sample collection and Identification: The whole plant *Beta vulgaris* was collected from a farm in Ooty, Nilgiris and *Psidium guajava* with the floral parts was collected from a farm in Devarayapuram, Coimbatore. The plants were authenticated by Scientist In-Charge, Botanical Survey of India, Tamil Nadu Agricultural University (TNAU), Coimbatore. The voucher specimens of the leaf samples (*Psidium guajava* - BSI/SRC/5/23/2020/Tech/555 and *Beta vulgaris* - BSI/SRC/5/23/2020/Tech/556) were stored in the herbarium of the department.

Cold percolation extraction: The leaf samples were dried under shade and ground to coarse powder. Cold extraction was performed using 100ml of ethanol for 10grams of the powdered leaf sample by providing a gentle shake continuously for around 24hours. This process was repeated for five batches and the extracts were collected, filtered using Whatman Grade 1 (11m) and stored in brown bottles for further experiments (Chiari-Andréo et al., 2017; Tripathy & Pradhan, 2013).

In vitro Scratch wound assay: The cell monolayer containing non- transfected L929 fibroblast cells was scraped using a 200µl pipette tip, washed to remove the debris and replaced with a suitable medium with the addition of the samples. They were then incubated at 37°C for around 18 hours and periodically checked for the cell-cell migration. The period of incubation allows the cell migration condition to complete the created scratch. Using, inverted phase contrast microscope, the scratch was captured before incubation and at regular time intervals, say, 4th, 18th and 24th hour when the positive control exhibited complete closure. Using, ImageJ software, the distance between 2 sides of the scratch i.e., the mean wound area (µm) during the healing process was measured (Liang, Park, & Guan, 2007; Santhini,et al , 2018).

RESULTS AND DISCUSSION

Sample collection and Extraction: The preparatory pretreatment methods for the collected leaves were carried out and extracted successfully using cold percolation method, which are essential for the extraction process. Based on the previous work in our lab (Udayakumar et al, 2020), the cold percolated ethanol leaf extracts of both *Beta vulgaris* and *Psidium guajava* were found to hold the phytochemical compounds, when compared with the other hot and cold solvent extracts of the leaf.

In vitro scratch wound healing activity: The in vitro scratch wound assay was performed on L929 Fibroblast cell lines (purchased from NCCS, Pune) cultured on minimum essential medium supplemented with Fetal Bovine Serum (FBS). The selected ethanol leaf extracts of beetroot and guava were added to the culture well plates in differing concentrations like 25µg, 50µg, 75µg and 100µg along with the standard positive control. The positive control was found to exhibit complete healing, where the wound area of 3835µm was completely closed at the 24th hour (Pooja et al, 2019) (Fig. 1a and Fig. 1b).

Beetroot showed increased cell proliferation on the created wound (wound area – 3331µm) at a minimum concentration of 25µg with the maximum wound closure of 78% after 24 hours. At increasing concentrations of, say, 50µg, 75µg and 100µg, they displayed lesser healing

activity due to the increased levels of ethanol suppressing the healing activity of beetroot. Ethanol exposure to the skin cells causes the delay in cell proliferation, collagen synthesis leading to slower epithelial coverage and blood vessel regrowth (Radek et al., 2005; Radek et al, 2009). Guava showed a lesser healing capacity on the created wound (wound area – 3205µm) when compared to beetroot with a healing potential of 45% for 25µg after 24 hours (Table 1).

Table 1. *In vitro* scratch wound healing activity of ethanol extract of extracts of Beta vulgaris and Psidium guajava under Cold percolation extraction

Sample	Concentration (µg)	Wound Area (µm)	Time intervals (h) and percentage of healing (%)			
			0	4	18	24
Beta vulgaris	25	3331	0	13	55	78
	50	3325		9	34	57
	75	3505		16	20	57
	100	3583		17	19	47
Psidium guajava	25	3205		36	38	45
	50	3453		32	37	42
	75	3559		31	41	42
	100	3739		30	31	41
Control	Unknown	3835		37	68	>99

Figure 1a: Microscopic images of in vitro scratch wound healing activity on L929 cells of ethanol extract of extracts of Beta vulgaris and under Cold percolation extraction

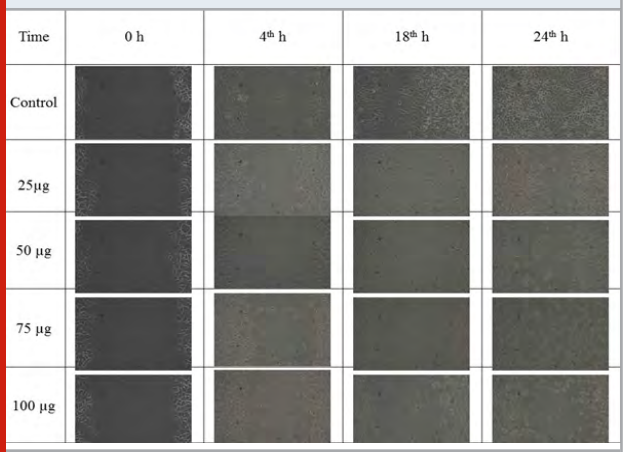


Figure 1b: Microscopic images of in vitro scratch wound healing activity on L929 cells of ethanol extract of extracts of Psidium guajava under Cold percolation extraction

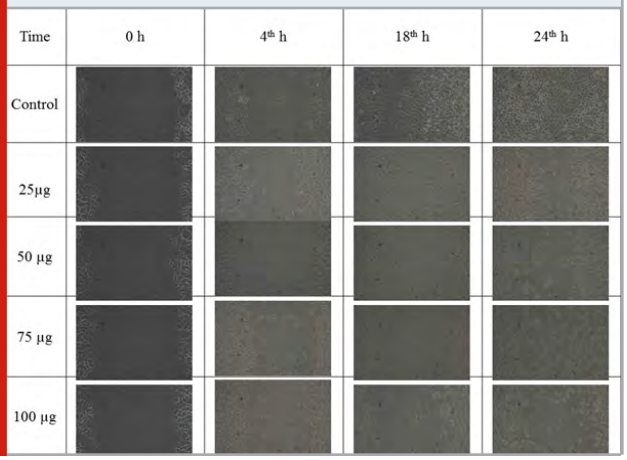
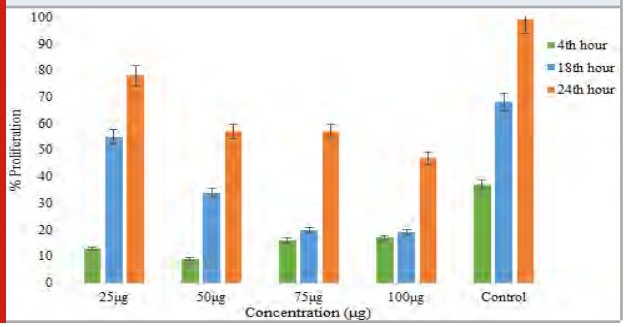


Figure 2a: Graph showing cell proliferation (%) of the ethanol leaf extracts of Beta vulgaris in differing concentrations captured at 4th, 18th and 24th hour time intervals



Similar to the beetroot extracts, guava also showed a decreasing linearity over its healing potential due to the inhibition of ethanol against epithelial cell proliferation (Fig. 2). Both the extracts showed an increasing continuance at their intervals, both, 4th and 18th hour with respect to their final healing activity at 24th hour. On the whole, both the selected plant extracts were found to exert comparable wound healing effects with the positive control (Porta et al., 2010).

CONCLUSION

The scientist fraternity is focusing on decoding the complex wound healing process to simplify the understanding of the process. Ethnopharmacological, conventional and synthetic methods of preparation of

formulations to treat wounds are existing. In addition to that phytochemical bio active compounds plays a significant role. The present study reported the wound healing activity of phytoconstituents extracts from *Beta vulgaris* and *Psidium guajava* and both the extracts exhibited comparable effects on treating wound. The research article would bring in new insights to concentrate on bio-prospect the plant sources to identify more potential bioactive molecules to encounter the global demand.

ACKNOWLEDGEMENTS

The authors are grateful to the Biotechnology Lab, SITRA, Coimbatore for assisting in the in vitro scratch wound healing analysis.

Conflicts of Interest: All the authors declare no conflict of interest.

REFERENCES

Chiari-Andréo, B. G., Trovatti, E., Marto, J., De Almeida-Cincotto, M. G. J., Melero, A., Corrêa, M. A., ... Isaac, V. L. B. (2017). Guava: Phytochemical composition of a potential source of antioxidants for cosmetic and/or dermatological applications. Brazilian Journal of Pharmaceutical Sciences, 53(2). https://doi.org/10.1590/s2175-97902017000216141

Gosain, A., & DiPietro, L. A. (2004). Aging and wound healing. World Journal of Surgery, 28(3), 321–326.

Guo, S. al, & DiPietro, L. A. (2010). Factors affecting wound healing. Journal of Dental Research, 89(3), 219–229.

Liang, C. C., Park, A. Y., & Guan, J. L. (2007). *In vitro* scratch assay: A convenient and inexpensive method for analysis of cell migration in vitro. Nature Protocols, 2(2), 329–333. https://doi.org/10.1038/nprot.2007.30

Liang, Z.-H., Pan, Y.-C., Lin, S.-S., Qiu, Z.-Y., & Zhang, Z. (2020). LncRNA MALAT1 promotes wound healing via regulating miR-141-3p/ZNF217 axis. Regenerative Therapy, 15, 202–209.

Pooja, R., Vadodaria, K., & Vidhya, S. (2019). Synthesis of bacterial cellulose and herbal extract for the development of wound dressing. Materials Today: Proceedings, 15, 284–293. https://doi.org/10.1016/j.matpr.2019.05.007

Porta, K., Fernandes, S., Bussadori, S. K., Marques, M. M., Sumie, N., Wadt, Y., ... Martins, M. D. (2010). Healing and cytotoxic effects of *Psidium guajava* (Myrtaceae) leaf extracts, 9(4), 9–14. https://doi.org/https://doi.org/10.20396/bjos.v9i4.8641730

Radek, K. A., Matthies, A. M., Burns, A. L., Heinrich, S. A., Kovacs, E. J., & DiPietro, L. A. (2005). Acute ethanol exposure impairs angiogenesis and the proliferative phase of wound healing. American Journal of Physiology - Heart and Circulatory Physiology, 289(3 58-3), 1084–1090. https://doi.org/10.1152/ajpheart.00080.2005

Radek, K. A., Ranzer, M. J., & DiPietro, L. A. (2009). Brewing complications: the effect of acute ethanol exposure on wound healing. Journal of Leukocyte Biology, 86(5), 1125–1134. https://doi.org/10.1189/jlb.0209103

Santhini, E., Pramila, V. M., Shalini, M., Vignesh Balaji, R., & Chellamani, K. P. (2018). Preparation and characterization of PLGA-based biocompatible nanoparticles for sustained delivery of growth factor for wound healing applications. Current Science, 115(7), 1287–1296. https://doi.org/10.18520/cs/v115/i7/1287-1296

Sorg, H., Tilkorn, D. J., Hager, S., Hauser, J., & Mirastschijski, U. (2017). Skin wound healing: an update on the current knowledge and concepts. European Surgical Research, 58(1–2), 81–94.

Thangapazham, R. L., Sharad, S., & Maheshwari, R. K. (2016). Phytochemicals in Wound Healing. Advances in Wound Care, 5(5), 230–241. https://doi.org/10.1089/wound.2013.0505

Tripathy, G., & Pradhan, D. (2013). Evaluation of in-vitro anti-proliferative activity and in-vivo immunomodulatory activity of *Beta vulgaris*. Asian Journal of Pharmaceutical and Clinical Research, 6(SUPPL.1), 127–130.

Udayakumar, G. P., Inbaraj, A., Baskar, K. K., Muthuraman, S., & Ramakrishnan, B. (2020). Study on the in vitro antioxidant properties of selected traditional medicinal plants. Journal of Pharmacognosy and Phytochemistry, 9(4), 1831–1837. https://doi.org/10.22271/phyto.2020.v9.i4y.12023

Biosc.Biotech.Res.Comm. Special Issue Vol 13 No 11 (2020) Pp-70-72

BBRC
Bioscience Biotechnology Research Communications

Medical Images Processing using Effectiveness of Walsh Function

Tamilarasu Viswanathan[1], M. Mathan Kumar[2] and C. Sasikumar[3]
[1,2,3]*Electrical and Electronics Engineering, Kumaraguru College of Technology, Coimbatore, India*

ABSTRACT

This paper presents the new method for processing medical images using effectiveness Walsh function. The Block pulse functions is defined, and the coefficient function is developed for identifying fixations and boundary limitations. The algorithms developed for basic functions with respect individual blocks and simulating using MATLAB. Proposed scheme shows that the performance analysis is better than existing schemes.

KEY WORDS: WALSH FUNCTION, BLOCK PULSE FUNCTIONS, FUNCTION COEFFICIENTS AND FIXATIONS.

INTRODUCTION

Many techniques employed for analyzing the medical images which are not clearly indicates the solution of the specific problem. This results in development of new methods are find out to identify advantages in specific method. The individual details are analyzing as functions of unique details is identified in simple and efficient way. Many repetitive methods are involved in set of well-defined functions means of trigonometric relationship and they are only suitable in limited boundary. It is necessary to propose a fresh method for wide range of solutions. Walsh function is suitable for medical image analysis and extend for the more independent analysis is proposed in this work. The objective of the medical image representation is adjustable in frequency domain for more set of function for doing the repetitive solutions. The image analyzing comprehensively used in solving nondeterministic problem for wide range including the extensive properties of relating many of them. More development occurs in computational effort using hardware and software arrangement leads to solve new category of problem statements.

Generated medical test images for specific case is shown in Fig. 1. From the regular analysis, the variation in the images giving different opinion about the medical report varies with person. In general, its necessary to give detail about the small variation make huge impact on the consultation. Including the Walsh function for analyzing the functions results in better analysis and use of effectiveness is discussed. In the next section, the Walsh function and its fundamentals are examined in the view of image analysis. In section III, the image parameters are related with the Walsh function is explained. The last conclusion section details the future scope and further improvement is discussed.

Walsh Functions and its Fundamentals: A. Block Pulse Functions (BPFs) The set of block pulses represented is shown in Fig.2 for every ith element N count is

$$W_i(t)=1, \text{ for } iT/N \le t \le (i+1)T/N \quad (1)$$

Where N is operating duration of individual pulse in seconds.

ARTICLE INFORMATION

*Corresponding Author: viswanathan.t.eee@kct.ac.in
Received 7th Oct 2020 Accepted after revision 13th Dec 2020
Print ISSN: 0974-6455 Online ISSN: 2321-4007 CODEN: BBRCBA

Thomson Reuters ISI Web of Science Clarivate Analytics USA and Crossref Indexed Journal

NAAS Journal Score 2020 (4.31) SJIF: 2020 (7.728)

*Online Contents Available at: http//www.bbrc.in/
Doi: http://dx.doi.org/10.21786/bbrc/13.11/16*

For every set, the individual components are identified through selection of operating limit. The general

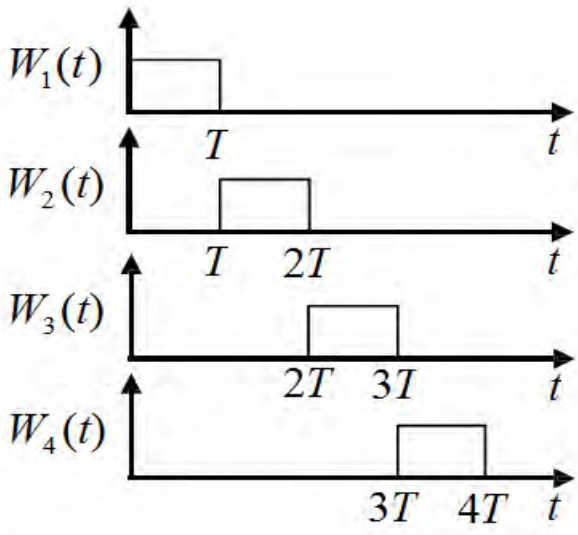
function can be set as operating limit for 1 to $\sqrt{N}$. Consider the function for doing image analysis as

$$f(t)=\sum_{i=0}^{\infty} C_i W_i(t)=C_1W_1(t)+C_2W_2(t)+\mathrm{L}$$
$$+C_NW_N(t)+\mathrm{L} \quad (2)$$

B. Representation of different combination of BPFs: The arrangement of any function is identified as individual BPF coefficient as (C1, C2...CN). Every individual sample is identifying the same N and T value for the solutions. The similarity rate identifying the flowchart shown in Fig. 3.

Figure 1: Medical Test Scan Image

Figure 2: General Block pulse Functions with Four Weights.



C. Fixation and Function coefficients: From the block boundaries, the limitations determined through a general form of Walsh function coefficients. They represent and explain the following conditions

1. The value should be regulated with image boundary and sample variations.
2. Comparing with existing schemes, the computational capability correctness with minimal approach.
3. The fixation and boundary to be verified for each case through necessary influences.
4. The intermediate transformations with providing the symmetrical and space variation.
5. Individual block should be combination of other adjacent blocks with general mentioning.

Figure 3

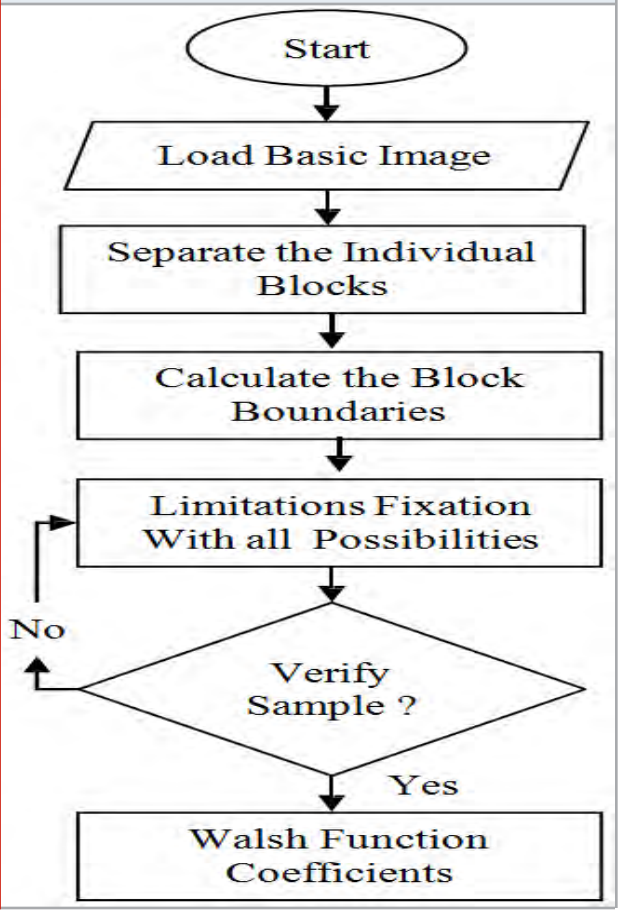


Figure 4: Walsh Transformation of Scanned image in Fig.1

RESULTS AND DISCUSSION

The proposed scheme is developed in MATLAB using m file from the flowchart shown in Fig. 3. The comparison

with existing scheme with proposed schemes is tabulated in Table 1.

Table 1. Comparision of Existing Scheme With Proposed Scheme

Properties	Existing Method [2-3]	Existing Method [3-4]	Proposed Scheme
Block separation	Not Available	Available	Available
Limitations	Pixel Missing	Computation Effort	File Size
Fixations	Yes	No	Yes
Accuracy	67%	85%	98 %
Depth Level	Level 2	Level 2	Level 3
Variation	Not Available	Maximum	Maximum
Determination	56 %	67-78 %	70-88 %

CONCLUSION

The effectiveness of Walsh function is used for identifying the features available in medical image for consultation. The images are analyzed with MATLAB and developed a generalized function for any level. In the future, the proposed method implements with hardware for real-time.

REFERENCES

Alireza Keshavarz (2020). walsh(N) (https://www.mathworks.com/matlabcentral/fileexchange/50202-walsh-n), MATLAB Central File Exchange

Arshad, U., Batool, S. I., & Amin, M. (2019). A novel image encryption scheme based on Walsh compressed quantum spinning chaotic Lorenz system. International Journal of Theoretical Physics, 58(10), 3565-3588.

ILambin, P., Leijenaar, R. T., Deist, T. M., Peerlings, J., De Jong, E. E., Van Timmeren, J., ... & van Wijk, Y. (2017). Radiomics: the bridge between medical imaging and personalized medicine. Nature reviews Clinical oncology, 14(12), 749-762..

Sneha, P. S., Sankar, S., & Kumar, A. S. (2020). A chaotic colour image encryption scheme combining Walsh–Hadamard transform and Arnold–Tent maps. Journal of Ambient Intelligence and Humanized Computing, 11(3), 1289-1308.

www.ucsfhealth.org/-/media/project/ucsf/ucsf-health/medical-tests/hero/knee-mri-scan-2x.jpg

Biosc.Biotech.Res.Comm. Special Issue Vol 13 No 11 (2020) Pp-73-78

BBRC
Bioscience Biotechnology Research Communications

Encapsulation and Characterization of Fucoidan-Curcumin Nano Micelle for Anti-inflammatory Effects

Balaji Sadhasivam and Saraswathy Nachimuthu*
[1]*Department of Biotechnology, Kumaraguru College of Technology, Coimbatore-641049, Tamilnadu, India.*

ABSTRACT

Curcumin is a potential bioactive compound used for many healthcare applications. Due to its hydrophobic nature therapeutic index is reduced many folds. Encapsulation of curcumin in a hydrophilic polymeric micelle enhances its bioavailability in biological systems. Fucoidan is a functional polysaccharide isolated from brown sea-weeds possess various health beneficial applications. In the present work, curcumin was encapsulated in fucoidan polymeric nano micelle (FCN). The optimal contact time for the encapsulation of curcumin was found to be 24 hours, average size of nano micelle formed 121 + 0.1 nm with 67% curcumin encapsulation efficiency. Zeta potential was maximum of -47.4 which shows electro-kinetic potential in colloidal dispersions of Fucoidan Curcumin Nano Micelle (FCN). Physical Characterization of FCN was performed using FTIR spectroscopy and conjugation of fucoidan and curcumin was confirmed based on peak shift at wave number 2215 cm^{-1}; 1639 cm-1; 1517 cm^{-1} and 1265 cm^{-1}. Analysis of SEM, and TEM images were performed to know the nature of encapsulation of curcumin in polymeric micelle. Results obtained from in vitro assays such as antioxidant, anti-hemolytic and anti-inflammatory showed potential bioactive properties of FCNs.

KEY WORDS: ANTI-INFLAMMATORY EFFECTS, FUCOIDAN-CURCUMIN, NANO MICELLE, TEM, SEM, FTIR.

INTRODUCTION

Curcumin has been used from earlier days as a remedy for different diseases. It shows good antioxidant activity by scavenging the reactive oxygen species (ROS) and nitrogen species. It is also proved to be a potent anti-inflammatory compound by altering several metabolic pathways mainly involving downregulation of key enzymes like cyclooxygenase-2 (COX-2) and 5-lipooxygenase (Rahmani et.al., 2018). It also has anti-microbial, anti-cancer, anti-diabetic and anti-mutagenic activity (Rahmani et.al., 2018; Gonzalez-Ortega et.al., 2020). Curcumin has a low bioavailability as it is hydrophobic in nature. Under in vivo condition, curcumin is quickly metabolized leading to poor absorption (Wal et.al., 2019; Phan et.al., 2018).

As a result, only a small amount of active biomolecule is available in the systemic circulation to exert therapeutic effects. Bioavailability of the curcumin could be increased by making a complex with a suitable hydrophilic molecule. Many methods are developed to increase the bioavailability of curcumin (Chen et al., 2020; De Leo et al., 2018; Karthikeyan et al., 2020; Moballegh Nasery et al., 2020; Wong et al., 2019; Youssouf et al., 2019). One of the potential methods by conjugating curcumin to polymers namely polysaccharides to make it hydrophilic. Natural polysaccharides are potential drug carriers as they are hydrophilic, biodegradable, biocompatible and can be modified (Naveen and Shastri, 2019).

ARTICLE INFORMATION

*Corresponding Author: saraswathy.n.bt@kct.ac.in
Received 9th Oct 2020 Accepted after revision 13th Dec 2020
Print ISSN: 0974-6455 Online ISSN: 2321-4007 CODEN: BBRCBA

Thomson Reuters ISI Web of Science Clarivate Analytics USA and Crossref Indexed Journal

NAAS Journal Score 2020 (4.31) SJIF: 2020 (7.728)

Online Contents Available at: http//www.bbrc.in/
Doi: http://dx.doi.org/10.21786/bbrc/13.11/17

Polysaccharides modified with hydrophobic molecules can improve its self-assembly and enabling them to form micelles with increased oral and topical absorption (Naveen and Shastri, 2019). In micellar structure the hydrophobic regions face inside while the hydrophilic regions are exposed to the aqueous environment (Tran and Tran, 2019).

Fucoidan is a natural sulphated polysaccharide with high fructose content of around 60% and hence it is widely used as health supplement for various applications. It is usually found in brown algae and is the most potent bioactive molecule. Studies have shown that fucoidan show good anti-inflammatory activity by inhibiting key enzymes which are responsible for inflammation. It is also a good antioxidant agent reducing the ROS and lipid peroxidation. Fucoidan is found to be a potential nanocarrier and has the ability to promote controlled drug delivery (Etman et.al., 2020). Studies have also shown that fucoidan to possess anti-inflammatory properties with suppression of enzymes like COX-2 (Manikandan et.al., 2020, Jeyawardena et.al., 2019). Conjugation of natural polysaccharides including pullulan and fucoidan with hydrophobic drugs and its use for cancer therapy has been previously studied (Grigoras, 2019, Guo et.al., 2020, Phan et.al., 2018). Fucoidan is a polysaccharide abundantly available in nature. Therefore, combining curcumin to fucoidan can improve the bioavailability of curcumin to increase its therapeutic effect.

In the current study, curcumin was encapsulated in fucoidan to form a Fucoidan Curcumin Nano micelle (FCN) and characterized using various techniques such as FTIR, SEM, TEM and zeta-potential. The bioactive properties of FCN were assessed by in vitro anti-oxidant, anti-haemolytic and anti-inflammatory assays.

MATERIAL AND METHODS

Materials: Fucoidan was extracted and purified from Sargassam longifolium a brown seaweed of Mandapam, Indian coastal region. Curcumin was procured from Sigma Aldrich, analytical grade ethanol and all other chemicals were used.

Preparation and Optimization Nano micelle: Fucoidan-curcumin nanoparticles were synthesized by following desolvation method (Maghsoudi et al., 2017). The known concentration of curcumin was prepared in desolvating agent absolute ethanol and the polysaccharide solution of fucoidan (1mg/mL) was dissolved in deionized water with 0.05% (v/v) Tween 20 as an emulsifying agent. The curcumin-loaded fucoidan nanoparticles were prepared by adding desolvating agent dropwise to the fucoidan solution under continuous stirring. After different contact time, contents were rapidly centrifuged at 10,000 rpm. The supernatant was centrifuged at 15,000 rpm for 15 to 20 minutes to precipitate the nanoparticle. The free curcumin from precipitate was removed by absolute ethanol wash. The synthesized nanoparticles were re-suspended in deionized water The curcumin encapsulation in nano-micelle were calculated by following formula (Kamaraj et al., 2018).

$$\text{Curcumin encapsulation (\%)} = \frac{\text{Amount of curcumin encapsulated}}{\text{Amount of Curcumin used}} \times 100$$

Characterization: Particle size of nano micelle was measured by a nano-zetasizer (Malvern, UK). Fourier Transform Infra-Red Spectroscopy FT-IR ATR (IRAffinity 1S, SHIMADZU) was used to study the changes in the molecular level in spectral range from 4000 to 450 cm^{-1} with a resolution of 1.0 cm^{-1}. Transmission electron micrograph (TEM) examination was performed in a Morgagni 268 D from FEI. A scanning electron micrograph (SEM) was recorded on a Jeol JSM 6390 microscope.

in vitro antioxidant assay: The free radical scavenging capacity of the nano micelle was evaluated by Diphenly picrylhydrazyl (DPPH) as free radical source (Chew et al., 2008). 20, 40, 60, 80 and 100 µL of the compound was taken in a tube and made up the volume to 3 ml with 0.1mM methanolic DPPH and incubated at 37˚C for 20 min. The absorbance was measured at 517 nm with methanolic DPPH solution as a negative control. The radical scavenging activity (RSA) was calculated as a percentage of DPPH discoloration using the equation:

$$DPPH\ Radical\ Scavenging\ Activity\ (\%) = \frac{OD\ of\ Control - OD\ of\ Test}{OD\ of\ Control} \times 100$$

in vitro anti-inflammatory activity-Human Red Blood Cells Membrane Stabilization assay: Anti-inflammatory assay was performed by investigating Fucoidan nano micelle FCN's human RBC membrane stabilization potential. The experiment was designed with modification of Bouhlali et al., (2016) method. The healthy volunteers with clear medical history from previous two weeks prior to experiment was selected. The whole blood was collected with equal volume of sterilized Alsever's solution (2% dextrose, 0.80% sodium citrate, 0.05% citric acid and 0.42% sodium chloride). Centrifuged at 2500 rpm for 10 minutes. The plasma was removed and cell suspension layer washed with isosaline (0.85% sodium chloride, pH7.2) and Centrifuged repeatedly until the supernatant was clear and colorless.

The HRBC component was resuspended in isosaline. The various concentrations of sample were prepared in distilled water. The reaction mixture equal volumes of sample, phosphate buffer (0.15M, pH7.4), hyposaline (0.36% sodium chloride) and HRBC suspension were incubated at 37˚C for 30 min. The content was centrifuged at 3000 rpm for 20 minutes and then the supernatant was observed at 560nm using spectrophotometer. Indomethacin was used as the reference drug. The percentage of membrane stabilization activity was calculated by using the formula, (Kumari et al., 2015).

$$\text{Percentage Stabilization} = \frac{\text{OD of Control} - \text{OD of Test}}{\text{OD of Control}} \times 100$$

in vitro anti-haemolytic assay: The haemolytic assay was performed by following method described by Mitra et al., (2015). Whenever RBC is subjected to irritating compound, cells will be lysed and haemoglobin will be released. The red blood cells were isolated from plasma and taken in Phosphate buffer saline. Different concentrations of samples were prepared in PBS. To the 950µl of sample add 50µl of RBC and incubate in dark for 10 minutes. Centrifuge the content at 6000rpm for 10 minutes to separate the cell debris. Measure the absorbance at 540nm. Consider deionized water as positive control and PBS as negative control.

$$\text{Percentage lysis} = \frac{\text{OD of Sample} - \text{OD of Positive Control}}{\text{OD of Negative Control} - \text{OD of Positive Control}} \times 100$$

RESULTS AND DISCUSSION

Particle Size and Zeta Potential Analysis: The fucoidan based polymeric micelle loaded with curcumin was synthesized by desolvating method. Fucoidan is L-fuco-polysaccharide consisting of number of fucose monomers with α(1,4) glycosidic linkage. Curcumin is a low molecular weight drug molecule under desolvating condition gets entrapped inside the polymeric chain forming nano micellar structure. Thus, obtained fucoidan nano micelles FCN were analyzed for particle size, zeta potential, curcumin encapsulation efficiency and yield.

The optimum contact time of polymer and drug molecule is important to form nano micelle harboring shell made of polymeric ring encapsulating drug molecule in its core (Karthikeyan etl., 2020). The table 1, shows the size of the particle and zeta potential value for different contact time and it is evident that the contact time affects the size of the particle, zeta potential and curcumin encapsulation efficiency. Increasing the contact time, size of the nano micelle formed decreased and at 48 hours of contact time, nano micelle fraction showed particle size diameter 140+ 0.1 nm and -46.9 zeta potential values with maximum curcumin encapsulation efficiency of 74.01 %.

The higher positive and lower negative zeta potential value ensures the stability and monodispersity of the nano particle (Basniwal et al., 2011). The maximum measured zeta potential value is -47.4 mV at 24 hours of contact time. This negative charge density of FCN implies that the hydrophobic drug curcumin encapsulated possess hydrophilicity and it is expected that it could result in increased bioavailability than the free curcumin in the biological system. Hence is concluded that 48 hours of contact time is optimum for nano micelle preparation.

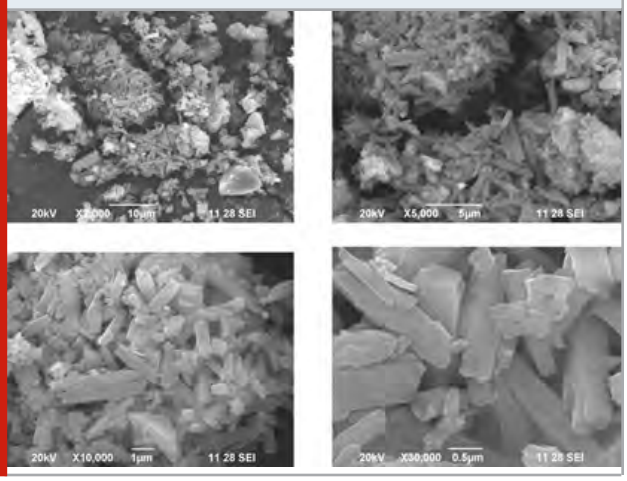
The surface topography of FCN nano micelle was examined using SEM as shown figure 1. The fucoidan-curcumin formed a distinct structure without any aggregation and irregular clusters. The FCN nano micelle is in size of range 121 nm to 196 nm. The figure 2(A &B) represents the TEM image of fucoidan-curcumin nano micelle and figure 2 (C, D &E) shows the core-shell properties of FCN analyzed using Image J software.

The figure 2 (D) shows the polymeric micelle having curcumin in its core and the figure 2(E) shows the fucoidan polymeric shell of FCN. Randomly 30 numbers of nano micelle structure were selected from processed TEM image of FCN (figure 2B) and analyzed for the core and shell diameter values. The Table 2 represents the average diameter and area of FCN. From the analysis, the average diameter of the core ranges from 84.7 nm to 160.9 nm and mean diameter is 115.6 + 18.3 nm. The average diameter of the shell ranges from 186.4 nm to 288.2 nm and mean diameter is 222.8 + 23.5 nm. This implies that fucoidan polymeric shell encapsulates the curcumin core as a nano micelle.

Table 1. Fucoidan curcumin Nano-Micelle at different Contact Time

Contact Time in (hrs)	Size of Particle (nm + SD)	Zeta Potential	Curcumin encapsulation (%+ SD)	Yield (%)
4	196+0.2	-19.5	63.11+0.12	45
12	153+0.1	-41.8	64.61+0.11	44
24	121+0.1	-47.4	66.92+0.13	49
36	132+0.2	-46.3	68.17+0.11	50
48	140+0.1	-46.9	74.01+0.21	50

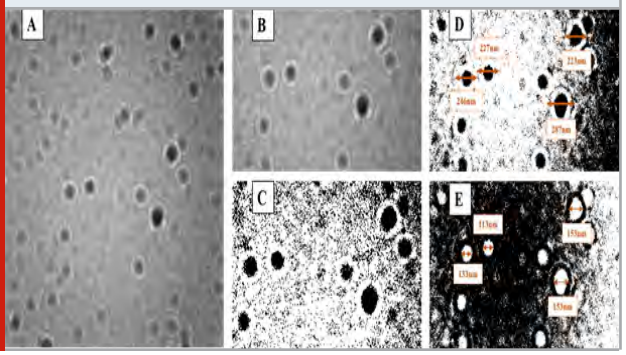
Figure 1: Scanning electron microscopic image of fucoidan-curcumin nano micelle



FT-IR Analysis: The Fourier Transform Infrared spectrum of fucoidan, curcumin and fucoidan curcumin nano micelle (FCN) is represented in figure. 3. The FTIR spectrum of curcumin, fucoidan and FCN at 665, 2100 and 3350 cm-1 wave number represents the -C=C-H:C-H bend of alkynes, -C= C- stretch of alkynes and O-H stretch vibrations of alcohols, phenols (Xu et al., 2014; Ye et al., 2013; Hifney et al., 2016). The fucoidan shows -C-N stretching of aromatic amines, CH stretching alkynes and C=C stretching vibrations of aromatic rings at 1269, 1388 and 1546 cm^{-1} respectively (Xu et al., 2014; Marudhupandi et al., 2015). The curcumin exhibited the presence of C-O Stretch alcohols carboxylic acids,

esters, ethers, -C-N stretching of aromatic amines and CH stretching alkynes at 1024 ,1200 and 2943 cm^{-1} respectively ((Mohan et al., 2012; Kamaraj et al., 2018).

Figure 2: Transmission Electron Microscopic image of Fucoidan Curcumin Nano Micelle with 500nm scale. (A) TEM image of FCN before processing (B-E) TEM image of FCN after processing in ImageJ software. (C, D) shows the core of the FCN with average diameter of 115.6 + 18.3 nm. (E) shows the shell of FCN with average diameter of 222.8 + 23.5nm



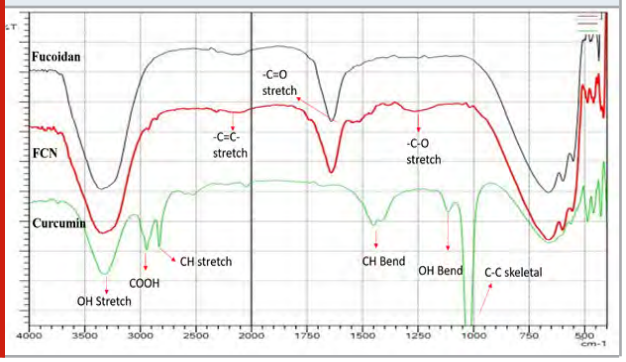
The wave number 1639 cm^{-1} and 2100 cm^{-1} shows the carbonyl C=O stretch of α, β unsaturated esters and presence of -C=C- stretches of alkynes of fucoidan was retained in FCN. The FCN shows peak shift at 1269 cm^{-1} from 1100 cm-1 of curcumin governing to -CO stretching vibrations of aromatic rings and another peak shift was observed from 1469 cm^{-1} to 1549 cm^{-1} of curcumin having -C=C stretching vibrations of aromatic rings. The COOH group, CH stretching of alkynes and C-C skeletal ring at 3311, 2943 and 1024 cm^{-1} of curcumin was not reflected in FCN indicating the encapsulation of curcumin in fucoidan polymeric micelle.

in vitro antioxidant assay: The DPPH Radical scavenging activity of curcumin, fucoidan and fucoidan-curcumin nano micelle was represented in figure 4. It was found that concentration of 100 µg/ml curcumin, fucoidan and fucoidan curcumin nano micelle (FCN) showed maximum scavenging activity of 98.9 %, 85.8 % and 97.1 % respectively. The IC_{50} value of FCN was 40.7 µg/ml which was similar to curcumin 39.1 µg/ml and lower than fucoidan 49.9 µg/ml. It was evident that increasing the concentration DPPH scavenging percentage was also increased and curcumin encapsulated in FCN enhanced the antioxidant potential.

Table 2. Fucoidan Curcumin Nano micelle Core and Shell diameter analysis from ImageJ software

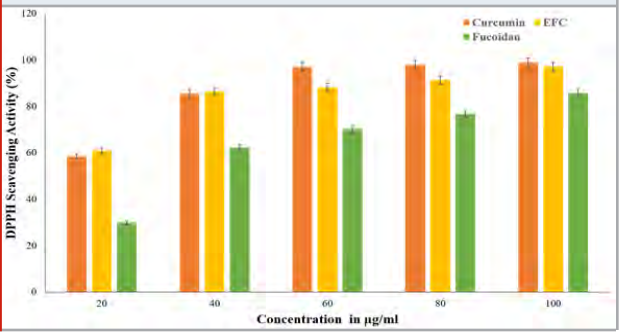
(in nm)	Nano micelle Area Range		Mean Nano micelle Area			Nano micelle Diameter Range micelle		Mean Nano Diameter
	Minimum	Maximum	Mean	SD	Minimum	Maximum	Mean	SD
Shell	1651.2	2512.7	1962.3	197.1	186.4	288.2	222.8	23.5
Core	789.7	1435.8	1050.5	153.7	84.7	160.9	115.6	18.3

Figure 3: FTIR Spectrum of Fucoidan Curcumin Nano Micelle (FCN), Curcumin and Fucoidan.



in vitro anti-inflammatory and anti haemolytic activity: The erythrocyte are predominantly available cells in human body and used extensively as biological model due to its morphological and physiological characteristics. During tissue injury or infection, the damaged cells and erythrocytes are exposed to the antigens leading to activation of immune responses. The stimuli activates inflammatory cells including macrophages, adipocytes, and the release of cytokines (TNF-α, IL-6 and IL-1β), inflammatory C-reactive protein (CRP) and enzymes such as glutathione peroxidase (GPx), high-mobility group box 1 (HMGB1), cyclooxygenase (COX)-2, inducible nitric oxide synthase (iNOS), superoxide dismutase (SOD) and NADPH oxidase (NOX), which independently restores homeostasis and controls the pathogen growth.

Figure 4: DPPH scavenging activity of Curcumin, Fucoidan and Fucoidan Curcumin Nano micelle



During abnormal inflammation, lysosomes releases excess enzymes into the cytosol which results in tissue damage. These inflammatory mediators target the erythrocyte membrane resulting in hemolysis and

haemoglobin oxidation (Chen et al., 2018). Non-steroidal anti-inflammatory drugs (NSAIDs) inhibits the release of the inflammatory mediator's or by stabilizing the lysosomal membrane rendering the anti-inflammatory effect (Mounnissamy et al., 2007).

Figure 5: *In vitro* **anti-inflammatory activity – hRBC membrane stabilization of Fucoidan Curcumin Nano micelle (FCN)**

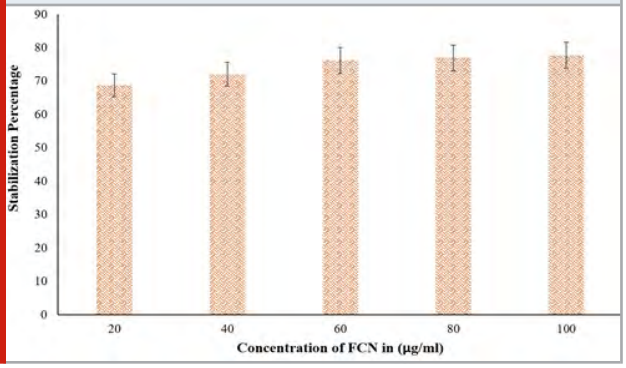
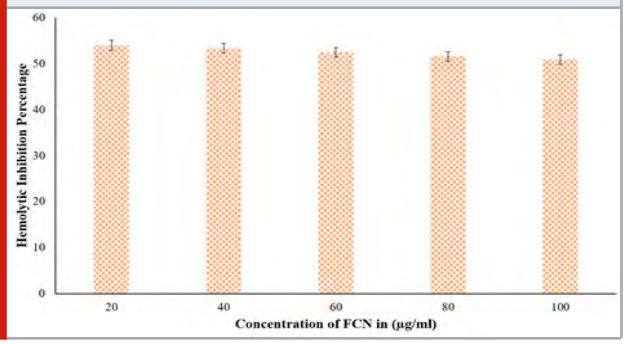


Figure 6: *In vitro* **anti-haemolytic activity – Hemolytic inhibition percentage of Fucoidan Curcumin Nano micelle (FCN)**



The human RBC stabilization activity of fucoidan curcumin nano micelle was represented in figure 5. It was found increasing the concentration hRBC membrane stabilization was also increased. The maximum stabilization was observed at 100µg/ml of FCN is 77% and IC_{50} was found to be 50.6µg/ml. The Anti hemolytic activity of fucoidan curcumin nano micelle was represented in figure 6. It was found that increasing the concentration fucoidan hemolytic inhibition was found to be the same at 54% and IC_{50} was found to be at 20.48µg/ml. Thus, the fucoidan curcumin nano micelle act as potential anti-inflammatory and blood compatible compound.

CONCLUSION

In this study curcumin was encapsulated in fucoidan to form a polymeric micelle (FCN). The successful encapsulation was confirmed by FTIR spectral analysis. SEM imaging showed no irregularities of nano clusters and no aggregates. TEM analysis showed the encapsulation of curcumin with curcumin in the core and fucoidan on the outer shell. Zeta potential showed the stability, monodispersity and hydrophilicity of FCN which signifies the potential for increased bioavailability of curcumin. The FCN showed to be a good antioxidant with 97.9% activity. It also showed to be a potent hRBC stabilizing agent and with good anti-hemolytic potential with IC_{50} of 20.48 µg/mL. The FCN has shown to possess several combined therapeutic benefits and with increased bioavailability serves as a potential dug candidate in various therapeutic applications.

ACKNOWLEDGEMENTS

This work was supported by Council of Scientific and Industrial Research, Ministry of Science and Technology, Government of India, India (Grant Number 08/677(0001)/EMR-I) and Kumaraguru College of Technology, India. The structural elucidation of fucoidan were done at Council of Scientific and Industrial Research Institute – Central Leather Research Institute, Chennai.

REFERENCES

Basniwal, R.K., Buttar, H.S., Jain, V.K. and Jain, N., (2011). Curcumin nanoparticles: preparation, characterization, and antimicrobial study. Journal of agricultural and food chemistry, 59(5), pp.2056-2061.

Bouhlali, E.T., Sellam, K., Bammou, M., Alem, C. and Filali-Zehzouti, Y., (2016). In vitro antioxidant and anti-inflammatory properties of selected Moroccan medicinal plants. Journal of Applied Pharmaceutical Science, 6(5), pp.156-162.

Chen, L., Deng, H., Cui, H., Fang, J., Zuo, Z., Deng, J., Li, Y., Wang, X. and Zhao, L., (2018). Inflammatory responses and inflammation-associated diseases in organs. Oncotarget, 9(6), p.7204.

Chen, Y., Lu, Y., Lee, R.J. and Xiang, G., (2020). Nano Encapsulated Curcumin: And Its Potential for Biomedical Applications. International Journal of Nanomedicine, 15, p.3099-3120.

Chew, Y.L., Lim, Y.Y., Omar, M. and Khoo, K.S., (2008). Antioxidant activity of three edible seaweeds from two areas in South East Asia. LWT-Food Science and Technology, 41(6), pp.1067-1072.

De Leo, V., Milano, F., Mancini, E., Comparelli, R., Giotta, L., Nacci, A., Longobardi, F., Garbetta, A., Agostiano, A. and Catucci, L., (2018). Encapsulation of curcumin-loaded liposomes for colonic drug delivery in a pH-responsive polymer cluster using a pH-driven and organic solvent-free process. Molecules, 23(4), p.739.

Etman, S.M., Elnaggar, Y.S. and Abdallah, O.Y., (2020). Fucoidan, a natural biopolymer in cancer combating: From edible algae to nanocarrier tailoring. International Journal of Biological Macromolecules, 147, pp.799-808.

González-Ortega, L.A., Acosta-Osorio, A.A., Grube-Pagola, P., Palmeros-Exsome, C., Cano-Sarmiento, C., García-Varela, R. and García, H.S., (2020). Anti-inflammatory Activity of Curcumin in Gel Carriers on Mice with Atrial Edema. Journal of Oleo Science, 69(2), pp.123-131.

Grigoras, A.G., (2019). Drug delivery systems using pullulan, a biocompatible polysaccharide produced by fungal fermentation of starch. Environmental Chemistry Letters, pp.1-15.

Guo, C., Hou, X., Liu, Y., Zhang, Y., Xu, H., Zhao, F. and Chen, D., (2020). Novel Chinese Angelica Polysaccharide Biomimetic Nanomedicine to Curcumin Delivery for Hepatocellular Carcinoma Treatment and Immunomodulatory Effect. Phytomedicine, p.153356.

Hifney, A.F., Fawzy, M.A., Abdel-Gawad, K.M. and Gomaa, M., (2016). Industrial optimization of fucoidan extraction from Sargassum sp. and its potential antioxidant and emulsifying activities. Food hydrocolloids, 54, pp.77-88.

Jayawardena, T.U., Fernando, I.S., Lee, W.W., Sanjeewa, K.A., Kim, H.S., Lee, D.S. and Jeon, Y.J., (2019). Isolation and purification of fucoidan fraction in Turbinaria ornata from the Maldives; Inflammation inhibitory potential under LPS stimulated conditions in in-vitro and in-vivo models. International journal of biological macromolecules, 131, pp.614-623.

Kamaraj, S., Palanisamy, U.M., Mohamed, M.S.B.K., Gangasalam, A., Maria, G.A. and Kandasamy, R., (2018). Curcumin drug delivery by vanillin-chitosan coated with calcium ferrite hybrid nanoparticles as carrier. European journal of pharmaceutical sciences, 116, pp.48-60.

Karthikeyan, A., Senthil, N. and Min, T., (2020). Nanocurcumin: A Promising Candidate for Therapeutic Applications. Frontiers in Pharmacology, 11.

Kumari, C.S., Yasmin, N., Hussain, M.R. and Babuselvam, M., (2015). Invitro anti-inflammatory and anti-arthritic property of Rhizopora mucronata leaves. Intern J Pharma Scie Res, 6, pp.482-485.

Maghsoudi, A., Yazdian, F., Shahmoradi, S., Ghaderi, L., Hemati, M. and Amoabediny, G., (2017). Curcumin-loaded polysaccharide nanoparticles: Optimization and anticariogenic activity against Streptococcus mutans. Materials Science and Engineering: C, 75, pp.1259-1267.

Manikandan, R., Parimalanandhini, D., Mahalakshmi, K., Beulaja, M., Arumugam, M., Janarthanan, S., Palanisamy, S., You, S. and Prabhu, N.M., (2020). Studies on isolation, characterization of fucoidan from brown algae Turbinaria decurrens and evaluation of it's in vivo and in vitro anti-inflammatory activities. International Journal of Biological Macromolecules, 160, pp.1263-1276.

Marudhupandi, T., Kumar, T.T.A., Lakshmanasenthil, S., Suja, G. and Vinothkumar, T., (2015). In vitro anticancer activity of fucoidan from Turbinaria conoides against A549 cell lines. International journal of biological macromolecules, 72, pp.919-923.

Mitra, T., Manna, P.J., Raja, S.T.K., Gnanamani, A. and Kundu, P.P., (2015). Curcumin loaded nano graphene oxide reinforced fish scale collagen–a 3D scaffold biomaterial for wound healing applications. RSC Advances, 5(119), pp.98653-98665.

Moballegh Nasery, M., Abadi, B., Poormoghadam, D., Zarrabi, A., Keyhanvar, P., Khanbabaei, H., Ashrafizadeh, M., Mohammadinejad, R., Tavakol, S. and Sethi, G., (2020). Curcumin Delivery Mediated by Bio-Based Nanoparticles: A Review. Molecules, 25(3), p.689.

Mohan, P.K., Sreelakshmi, G., Muraleedharan, C.V. and Joseph, R., (2012). Water soluble complexes of curcumin with cyclodextrins: Characterization by FT-Raman spectroscopy. Vibrational Spectroscopy, 62, pp.77-84.

Mounnissamy, V.M., Kavimani, S., Balu, V. and Quine, S.D., (2007). Evaluation of Anti-inflammatory and Membrane stabilizing property of Ethanol Extract of Cansjera rheedii J. Gmelin (Opiliaceae). Iranian Journal of Pharmacology and Therapeutics, 6(2), pp.235-0.

Naveen, C. and Shastri, N.R., (2019). Polysaccharide nanomicelles as drug carriers. In Polysaccharide Carriers for Drug Delivery (pp. 339-363). Woodhead Publishing.

Phan, N.H., Ly, T.T., Pham, M.N., Luu, T.D., Vo, T.V., Tran, P.H. and Tran, T.T., (2018). A comparison of fucoidan conjugated to paclitaxel and curcumin for the dual delivery of cancer therapeutic agents. Anti-Cancer Agents in Medicinal Chemistry (Formerly Current Medicinal Chemistry-Anti-Cancer Agents), 18(9), pp.1349-1355.

Rahmani, A.H., Alsahli, M.A., Aly, S.M., Khan, M.A. and Aldebasi, Y.H., (2018). Role of curcumin in disease prevention and treatment. Advanced biomedical research, 7.

Tran, T.T. and Tran, P.H., (2019). Nanoconjugation and encapsulation strategies for improving drug delivery and therapeutic efficacy of poorly water-soluble drugs. Pharmaceutics, 11(7), p.325.

Wal, P., Saraswat, N., Pal, R.S., Wal, A. and Chaubey, M., (2019). A Detailed Insight of the Anti-inflammatory Effects of Curcumin with the Assessment of Parameters, Sources of ROS and Associated Mechanisms. Open Medicine Journal, 6(1).

Wong, K.E., Ngai, S.C., Chan, K.G., Lee, L.H., Goh, B.H. and Chuah, L.H., (2019). Curcumin nanoformulations for colorectal cancer: a review. Frontiers in pharmacology, 10, p.152.

Xu, P., Wu, J., Zhang, Y., Chen, H. and Wang, Y., (2014). Physicochemical characterization of puerh tea polysaccharides and their antioxidant and α-glycosidase inhibition. Journal of Functional Foods, 6, pp.545-554.

Ye, H., Zhou, C., Li, W., Hu, B., Wang, X. and Zeng, X., (2013). Structural elucidation of polysaccharide fractions from brown seaweed Sargassum pallidum. Carbohydrate polymers, 97(2), pp.659-664.

Youssouf, L., Bhaw-Luximon, A., Diotel, N., Catan, A., Giraud, P., Gimié, F., Koshel, D., Casale, S., Bénard, S., Meneyrol, V. and Lallemand, L., (2019). Enhanced effects of curcumin encapsulated in polycaprolactone-grafted oligocarrageenan nanomicelles, a novel nanoparticle drug delivery system. Carbohydrate polymers, 217, pp.35-45.

Biosc.Biotech.Res.Comm. Special Issue Vol 13 No 11 (2020) Pp-79-81

BBRC
Bioscience Biotechnology Research Communications

Malicious URL Detection Using Rule Based Optimization Techniques

N. Jayakanthan[1]* and R.M. Anu Varshini[2]
[1]*Assistant Professor (SRG), Department of Computer Applications, Kumaraguru College of Technology,Coimbatore, India
[2]Department of Computer Applications, Kumaraguru College of Technology, Coimbatore, India

ABSTRACT
The suspicious URL cause harms to users dealing with online tractions. The malicious URLs are harmful to society and induvial user. It attacks the victims computer and steal all their confidential information. Such malicious URL to be analyzed identified and blocked. In this paper we prose rule based model DETECTX to detect the malicious URL. This algorithm analyses various features of the URL for classification. The experimental result shows the efficiency of the proposed system.

KEY WORDS: MALICIOUS URL, RULE BASED CLASSIFICATION, URL DETECTION, MALICIOUS, WEB PAGE.

INTRODUCTION

The taint URL attracts the user to visit the suspicious web site. It collects the details. Like users personal information. It spoils the users system to carry out the attack. These kind of attacks are very serious concern in present cyber security era. Hence the users should conscious about cyber attack. The existing approaches using profiles and machine learning. Even though they have some merits but they have major pitfalls also. Most of the current method identifies the traditional pitfalls. But lot of development in malicious URLs and new features are introduced. In United States a report says 15000 malicious attacks are carried out in every second. It is an imbalance situation. The cyber user to be protected from the malicious attacks. So there is need for research works in this area. The proposed approach is a based on Ant Colony Optimization. The taint features of the URL are analyzed and clustered. Based on the features the URL is classified as genuine or malicious. The experimental results show the efficiency of this approach.

Azeez chaudray et all introduced a system which analyzes the lexical attribute of the web page using linear algebra to identify the cyber crime. Babu Kannan et al produce report which analyzes the impact of various cyber vulnerabilities, He reports it will increase in future. N. Jayakanthan et al analyze the various features in malicious URL and identify concern URL is harmful or not. ID3 decision tree algorithm is used for this purpose. N. Jayakanthan et al uses the Graph based classifier to find the taint URL and categorize it into malicious or genuine. N.Jayakanthan et al carried out a research activity to find various malicious activates occurs in web applications. Both staic and dynamic models are explained.

Kzek ymi developed tool URL finder to identify the URL performing criminal activity by analyzing the properties of the URL. Mousavinejad et al identify various malicious

ARTICLE INFORMATION

*Corresponding Author: jayakanthan.n.mca@kct.ac.in
Received 20th Oct 2020 Accepted after revision 5th Dec 2020
Print ISSN: 0974-6455 Online ISSN: 2321-4007 CODEN: BBRCBA

Thomson Reuters ISI Web of Science Clarivate Analytics USA and Crossref Indexed Journal

NAAS Journal Score 2020 (4.31) SJIF: 2020 (7.728)

Online Contents Available at: http//www.bbrc.in/
Doi: http://dx.doi.org/10.21786/bbrc/13.11/18

activities in network control system. Recursive algorithm is used to detect the attack. Musoodjafran et alnarrate a from work which analyze the various factors to URL. To set the URL is malicious or genuine. This approach narrates the various attribute find the malicious URLs. Sheryas analyzes prose linear model to find malicious URLs which detects the cyber security violations in public gathering media.Ying Da et al analyze the URL pattern to segregate the given URL is malicious or not. The authors mine the URL patterns for the classification. The remaining section of the paper is organized as follows. Chapter 2 Material and Methods, the results are given in the chapter 3. Chapter 4 concludes the research work.

MATERIAL AND METHODS

Malicious URL detection is a imperative methodology for the present scenario. The online users are impacted by the malicious attacks. These website steals the user's personal information. Utilize the user's machine for further attack. Hence these attacks to be detected and users to be prevented from these attacks. Lot of research works are carried out in this area. But these methods are having major drawbacks. They are detecting traditional attacks but attackers are launching new attacks. Hence it is essential to carry out new research work in this area. In this research work we proposed rule based approach to detect malicious URL. It compares the occurrence of the various malicious features and detect the given URL is genuine or malicious. List of features are given the following table 1.

Table 1. Features of Malicious URL

SL.No	Features
1	Host name
2	Taint Special Character
3	Cap symbol (^)
4	Dot count (.)

Table 1 gives the features of malicious URL. The Host name belongs to malicious repository then the URL is declared as malicious. If URL contain a taint special character like @ character then it is malicious. If cap symbol occurs then it is declared as malicious. If dot count > 5 then the URL is suspicious. The algorithm DETECX of the proposed approach is given below:

Algorithm DETECX(URL)
Input : Uniform Resource Locator (URL)
Output : Taint or legal
T_R – Taint Feature set
Status – Variable store the result.
Step 1 : Set Status = Legal
Step 2 : Initialize T_R = NULL
Step 3: If Host name (HS) € Malicious Repository then
Set Status = Taint
$T_R = T_R \cup HS$
Step 4: If URL contain taint special character (Ts)
Set Status = Taint
$T_R = T_R \cup T_S$
Step 5: If URL contain cap symbol (^)
Set Status = Taint
$T_R = T_R \cup ^$
Step 6: If Dot Count (Dc) > 5 then
Set Status = Taint
$T_R = T_R \cup D_c$
If Status = Taint then Display the URL is Taint
Display List of malicious features
Else
Display the URL is legal

The proposed rule based algorithm analyze the features if any malicious activity found the URL is declared as malicious otherwise the URL is declared as genuine.

Figure 1: Performance Analysis

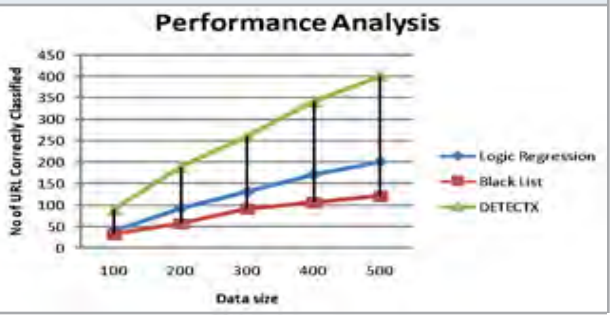


Table 2. Result Analysis

	Logic Regression	Black list	DETECTX
Number of URL Correctly Classified	200	120	400
Success ratio	40	24	80
Failure ratio	60	76	20

RESULTS

The proposed approach is implementing as Java class. It analyzes all features listed in table1. The malicious URL is collected from Phishtank and genuine URLs are collected from DOMZ. In total 500 URLs are collected 250 URLs are legal and 250 URLs are taint. 200 URLs are used for training and 300 URLs are used for testing. To check the efficiency of the proposed approach the testing and training set are kept as disjoint. The proposed approach is compared with other approaches using Logic Regression and Block list based approach the experimental results are given figure 1.

The experimental results shows our approach is efficiently classify the URLs than other approached the classification accuracy is shown in the following table2. The proposed approach is compared with logic regression and black listed based approaches. The success rate of the proposed

system is 80% which is higher than other approaches. It shows the efficiency of the proposed approach.

CONCLUSION

The cyber attacks are the significant threat to the society. It impact the user by accessing the personal information. It essential to prevent such attacks. In this paper a naval framework DETECTX is proposed to detect malicious URL. Rule based clustering algorithm is used for this purpose. Our approach is compared with Logic regression and Blacklist based models. The experimental results show the efficiency of the proposed approach.

REFERENCES

Azeez chaudray(2019) Cyber Crime identification, Ireland Journal of Cyber Law, pp.1021-1057.

Babu Kannan(2017) Impact of Security Vulnerability – An Analysis, Proceeding of 6th International Conclave, Delhi, India.

N.Jayakanthan and A.V.Ramani(2016) A Feature Based Framework to Detect Malicious URLs, International Journal of Control Theory and Applications pp.1327-1340.

N. Jayakanthan and A.V.Ramani(2017) Graph based Classifier to Detect Malicious URL, International Journal of Mechanical and Production Engineering Research and Development, pp. 223-234.

N.Jayakanthan and M.Manikantan, "Malicious Attack Detector(2017), International Journal of Advance Research and Innovative Ideas in Education.Issue .

Kzek ymi, and Nova berk(2017) URL Finder- An Impertive tools, Proceeding of the 7th International Conference on Computer Science, Amristsar, India.

Kumar.S (2014) Malicious Web Attacks Detectio", Journal of Cyber Security.

Mousavinejad,Fuwen Yang,Qing-Long Han and Ljubo Vlaci,A(2018) Novel Cyber Attack Detection Method in Networked Control Systems, IEEE Transactions on Cybernetics.

Musoodjafran,(2018) URL : is Indicator to identify cyber attack, Proceedings of International Conference on Computing Technology, Chennai, India, August.

Sheryas v(2018), Detecting Cyber Attack in Public Sites, International Journal of Technology.

Suji T(2019) Malicious URL Indication International Journal of Engineering.

Tiwan v and Suki S(2016) Phising URLs a Over view.

Ying Da, HuangKai XuJian Pei(2013)

Malicious URL detection by dynamically mining patterns without pre-defined elements, Journal of World Wide Web.

Zanvu and Miam(2015) Phishing Attack Classification, Journal of Computer Science.

Biosc.Biotech.Res.Comm. Special Issue Vol 13 No 11 (2020) Pp-82-87

BBRC
Bioscience Biotechnology Research Communications

Mobile Based Leaf Disease Classifier

Chandrakala D[1], Sarath Kishore R[2], Kishore R[3], Nandha Kumar M. K[4]
[1]*Professor, Department of Computer Science and Engineering,*
[2,3,4]*Student, Department of Computer Science and Engineering,*
Kumaraguru College of Technology, Coimbatore, Tamil Nadu , India

ABSTRACT

Crop diseases pose a significant threat to food security and yield, but their rapid and efficient identification is still a complicated and cumbersome process in many parts of the world because of the lack of related infrastructure. The advent of A.I., current advances in the areas of image processing, and the increasing dispersion of mobile devices into the masses, advocates the idea of mobile-based disease identification. Using a public dataset of over 20000 images of healthy and infected tomato, potato leaves that were collected under certain controlled conditions, a deep Convolutional Neural Network (CNN) has been trained to analyze and identify 15 diseases of the said leaves. The trained model achieved an accuracy of 89.325% on a test set, illustrating the feasibility of this method. Overall, this method of training diverse deep learning data models on progressively large and diverse image datasets lays a clear foundation for mobile-based crop disease identification on a massive scale.

KEY WORDS: DISEASE CLASSIFIER, CONVOLUTION NEURAL NETWORK, VANISHING GRADIENT, NORMALIZATION, POOLING.

INTRODUCTION

Human life has largely depended on edible crops to survive, for a long time. Modern technologies have enabled human society to produce enough food crops to meet the demand of the growing population of 7 billion and maintain sustainability. However, there is still a major threat to food security owing to several factors such as pest infestation, plant diseases, fall in pollination, climate changes. Plant diseases are a major threat not only for food sustainability but also for small scale farmers and other such people whose livelihood depends mainly on healthy crops and about 80% of the agricultural production is generated by the likes of such people.

ARTICLE INFORMATION

*****Corresponding Author: chandrakala.d.cse@kct.ac.in
Received 25th Oct 2020 Accepted after revision 9th Dec 2020
Print ISSN: 0974-6455 Online ISSN: 2321-4007 CODEN: BBRCBA

Thomson Reuters ISI Web of Science Clarivate Analytics USA and Crossref Indexed Journal

NAAS Journal Score 2020 (4.31) SJIF: 2020 (7.728)

Online Contents Available at: http//www.bbrc.in/
Doi: http://dx.doi.org/10.21786/bbrc/13.11/19

Various methods are put in place to prevent crop loss due to pests and diseases. Traditional methods include the widespread application of pesticides. But it is essential to identify the disease correctly when it appears in order to apply the appropriate pesticide. It is a crucial step in efficient disease management. Historically, this is done by agricultural organizations and universities and other such institutions wherein the same is sent to the concerned expert who analyses the sample and determines the route of remedy.

In recent times, with the advent of the Internet, such traditional efforts are additionally supported and by the instant provision of relevant information aiding to an online diagnosis process. There are billions of mobile devices in use among the masses today, further advocating the processes related to an online solution. Smartphones in particular offer a variety of new approaches with their efficient computing power, high-resolution cameras and displays, a wide array of sensors and accessories prebuilt into them. All these features packed together make it plausible for a situation where

image-based disease diagnosis based on neural networks can be made available on an unprecedented scale. Here, the feasibility of using the Convolutional Neural Network (CNN) approach utilizing 22,953 images of 3 crop species with 15 diseases and healthy leaves from the publicly available Plant Village dataset is established.

The Large-Scale Visual Recognition Challenge (ILSVRC) which is based on the ImageNet dataset has been widely used as benchmarks for numerous problems in computer vision including object recognition. Following the top-5 error rate of 16.4% achieved by a deep convolutional neural network for classification of 1000 categories, many advances have been made in the field of deep convolutional networks which has significantly reduced the error. Training of large neural networks on huge datasets is very time-consuming and requires heavy computing power. But once these models are trained, they can be deployed anywhere. The trained models can classify the images quickly and do not require much computing requirements, making it very much suitable for smartphones. To make accurate classifications, the model must be trained over a diverse and large dataset. This problem was solved by the Plant Village dataset project where they collected tens of thousands of images of healthy and diseased crops of various species. Here, a deep convolutional model that classifies 15 diseases in 3 crop species has been trained.

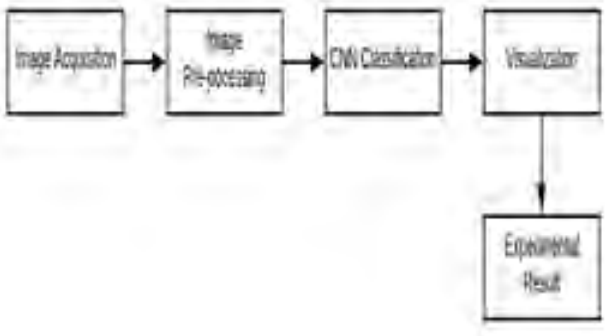
The best performing model achieved an accuracy of 89.325% thus advocating the feasibility of this approach. Vaijinath B et. al. implemented leaf disease detection using image processing and Support Vector Machine (SVM), the input image was first blur softened to reduce the noise (Vaijinath B Batule, Gaurav U Chavan, Vishal P Sanap, Kiran D Wadka, 2016). Then the image was converted from RGB (Red, Green, Blue) to HSV (Hue, Saturation, Intensity) for efficient color thresholding. Then separation of foreground and background was done for leaf segmentation from the image. After the image segmentation feature extraction was performed and given as input to the SVM and K-means algorithm to get the desired results.

Belal A M Ashqar and Samy S. Abu-Naser presented an Image-Based Tomato Leaves Diseases Detection Using Deep Learning. (Ashqar, Belal, Abu-Naser Samy, 2018). They used tomato leaf images of 5 different disease categories from the plant village dataset namely Bacterial Spot, Early Blight, Healthy, Septorial Leaf Spot, Leaf mold, Yellow Leaf Curl Virus. They implemented a Deep learning technique and arrived with an accuracy of 99.84% for full color and 95.54% for grayscale model. Drjoy Sen Maitra et. al. presented a CNN based approach to handwritten character recognition for multiple scripts of India. (Sen Maitra Durjoy, Bhattacharya, Ujjwal, Parui, Swapan. 2015).

They used 6 datasets of 100 classes namely Bangla basic characters, Bangla numerals, Devanagari numerals, Oriya numerals, Telugu numerals, English (MNIST - Modified National Institute of Standards and Technology database) numerals. The images were converted into grayscale and trained for more than 1000 iteration to get an accuracy of 85% and used the obtained feature as an input to the SVM model and the desired output was produced where the accuracy of Bangla basic characters, Bangla numerals, Devanagari numerals, Oriya numerals, Telugu numerals, English (MNIST) numerals are 95.6%, 98.37%, 98.54, 97.2%, 96.5%, and 99.10% respectively.

Sabah Bashir and Navdeep Sharma presented an image processing technique to detect diseases in Malus Domestica using image processing in MATLAB (Sabah Bashir, Navdeep Sharma, 2012). The image was taken in RGB format to facilitate the creation of a histogram for the comparison of images. K-means clustering technique was used for the classification of images based on disease. Then threshold separation of the image was done to separate the background and foreground. Sujatha R et. al. presented a technique to identify diseases in the leaf efficiently (Sujatha R, Sravan Kumar Y, Garine Uma Akhil 2017). Using the K-means clustering image was segmented into K groups representing segments of the image and SVM technique used to distinguish the disease and hence arrive at the desired results. Prajakta Mitkal et. al. presented an image processing technique for the detection of sugarcane leaf diseases (Prajakta Mitkal, Priyanka Pawar, Mira Nagane, Priyanka Bhosale, Mira Padwal and Priti Nagane, 2016). Authors used SVM, Non-Linear SVM, and Multiclass SVM for feature extractions and obtained high accuracy.

Figure 1: Proposed Methodology



Prakash M Manikar et. al. described a plant disease recognition technique which uses the Spatial Gray-level Dependence Matrices (SGDM) and K-means clustering (Prakash M. Mainkar, Shreekant Ghorpade,, Mayur Adawadkar, 2015). K-means technique is used to segment the image. The infected clusters are converted from RGB color space to HSI and the SGDM matrix was generated for Hue and Saturation. In the next phase, the GLCM function was used to calculate the features and compute the texture statistics. In the last phase, extracted features were passed through a neural network that is trained for disease recognition. Ghulam Mustafa Choudhary and Vikrant Gulati presented an image processing technique for the detection of Scorch and Spot diseases of plants (Ghulam Mustafa Choudhary, Vikrant Gulati, 2015). Firstly, the leaf image is acquired in the RGB color format.

Then a color transformation structure was created and the RGB color values are converted to that space in the pre-processing. K-means clustering was applied, and green pixels were masked to remove those cells inside the boundaries of infected clusters and obtain useful segments of the image. The color co-occurrence method was used to compute color, texture, and edge features. Lastly, the disease was classified by a configured neural network.

Hrishikesh P. Kanjalkar and S.S.Lokhande explained how feature extraction is useful in simplifying the computation power and memory size due to a combination of variables (Hrishikesh P. Kanjalkar, S.S.Lokhande, 2014). In this work, the leaf spots are segmented from the leaves using the application of graph theory algorithm. Pradnya Ravindra Narvekar et. al. identified the disease in the grape leaf by making use of image processing techniques (Pradnya Ravindra Narvekar, Mahesh Manik Kumbhar, S N Patil, 2014). Distance between the concurrent spots was identified and the distance was calculated. Features considered in this method are color and texture of the image. This method calculates two values such as cluster shade and cluster prominence, by comparing these values the disease can be identified. Vijai Singh and Mishra detected a plant infection using image processing techniques (Vijai Singh,, A. K. Mishra 2017). The infections considered here are Anthracnose, Cercospora leaf spot, and bacterial blight. The images were either captured from the mobile or retrieved from mobile storage. From the texture values obtained from the Gray Level Co-occurrence Matrix (GLCM), the disease was classified. Ghulam Mustafa Choudhary extracted features such as color, texture, morphology, and structure by using the GLCM method. The model was built using Support Vector Machine to classify the diseases.

Sharada Prasanna Mohanty experimented deep learning for the detection of plant disease (Sharada Prasanna Mohanty 2016). The deep learning architecture involved were AlexNet and GoogLeNet. The dataset involved in the study contained 54,306 consisting of 38 classes (14 crop species and 26 diseases). For interpreting the model with more accuracy, varying configurations were used by changing the color space representation and training and testing set distribution. The color space representation consisted of color, grayscale, and leaf segmented, and the training-testing distribution as 80%-20%, 60%-40%, and 50%-50%. The highest accuracy level obtained by varying the configuration was 98.36%, the configurations used were GoogLeNet, Color and 80%-20% distribution ratio.

METHODOLOGY

Convolution neural networks is a part of deep neural networks, which makes the algorithm to precisely identify the deeper level of the image as a composition of edges, curves, line, and other contents, instead of just going through the image roughly like in other neural networks. CNN is inspired by the visual cortex of a human brain, which has certain regions of cells specific to certain regions of the visual field. Features of the image(filters) that have to be manually fed to classify an image in traditional image classification techniques were made to learn by itself on CNN so it drastically reduced the human effort of feeding it. An input of image is passed through hidden layers namely convolutional, nonlinear, pooling (downsampling), and fully connected layers, and get an output which can be a single class or a probability of classes that best describes the image. The purpose of this layer is to extract the features of the image that describe the input image and it is the first layer of CNN. Here a filter size is defined which is based on, how detailed the precision of feature recognition, smaller the filter matrix higher the precision.

A filter is taken and multiplied with the image matrix from the top and traversed column wise followed by rows. As a result, a feature map that identifies the curves, edges, lines, or features of the image is obtained. The region of the image matrix that is multiplied with is called the receptive field. Padding is used to augment the shrunken feature map to the original size of the image or to make it bigger than the original image. This must be done to avoid getting a reduced feature map that is 2x2 matrix after each convolution layer. Activation layer essentially follows each convolution layer. When the system is working on linear operations during Conv layers, activation function is added to bring non-linearity to the function.

Vanishing gradient problem is the issue where the gradient decreases exponentially from layer to layer, which also causes the lower layer to train slowly and this issue is solved by ReLU. The function $f(x) = \max(0,x)$ is applied to all the ReLU layers to all of the values from the input to this layer. ReLU turns all the negative values to zero. This layer increases the nonlinear properties of the model and the overall network without affecting the receptive fields of the conv layer.

Pooling is used to reduce the image size in terms of width and height. It is of two types, max-pooling which reduces the image size by taking the largest value in that window and average pooling which reduces the image size by taking the average value in that window. Dropout is used to drop a random activation function by setting it to zero which prevents the model from being overfitted to the training set. The model should even be able to identify and classify images based on feature. It is used to convert the pooled feature map into a sequential column of numbers like a long vector. Then this is used as an input layer of an artificial neural network for further processing. The first step is to acquire images of various layers CNN is used to extract features and classify them based on the feature map. Figure 1 shows the flow of the process of proposed methodology. The flow of the process of proposed system is: (i) Image Acquisition (ii) Image Preprocessing (iii) CNN Classifier (iv) Visualization (v) Experimental Result.

Image Acquisition: The dataset is obtained from the public Plant Village project database, which contains

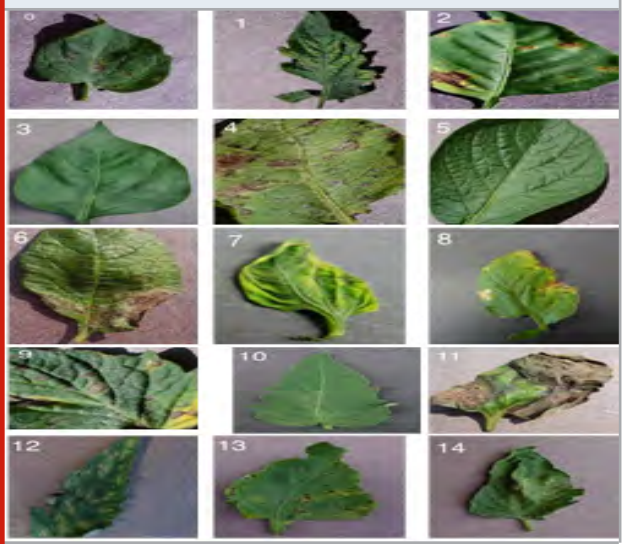
22,953 images of 3 crop species and 12 diseases. This database consists of 15309 images of Tomato leaves, 4668 images of Potato leaves, and 2974 images of Pepper Bell. This dataset has samples of healthy leaves in addition to diseased leaves. https://ww.kaggle.com/emmarex/plantdisease

It has 15 classes in total as follows:

- Class(0): Tomato Target Spot`
- Class(1): Tomato Mosaic Virus
- Class(2): Pepper Bell Bacterial Spot
- Class(3): Pepper Bell Healthy
- Class(4): Potato Early Blight
- Class(5): Potato Healthy
- Class(6): Potato Late Blight
- Class(7): Tomato Yellow Leaf Curl Virus
- Class(8): Tomato Bacterial Spot
- Class(9): Tomato Early Blight
- Class(10): Tomato Healthy
- Class(11): Tomato Late Blight
- Class(12): Tomato Leaf Mold
- Class(13): Tomato Septoria Leaf Spot
- Class(14): Tomato Spider Mites

Image Preprocessing: The acquired images are converted into a 256x256 image size. Due to differences in a number of images in each class of the dataset, image augmentations like rotation, width shift, height shift, brightness, shearing, horizontal flip, vertical flip are used to generate new images from existing datasets to maintain a standard number of images in each class, To achieve this ImageDataGenerator from Keras was used. Then the images are divided based on the train and test ratio and passed into the CNN classifier.

Figure 2: Dataset Samples



Classification: This model takes color images as input, so Convolutional Neural Networks (CNNs) are used to extract features from the leaf images. The resulting model has two parts: The first part of the model has Convolution layers with ReLU activation function, Normalization layers, Max Pooling layers, and Dropout layers arranged in sets of different combinations. The second part of the model is the set of Dense and Flatten layers combined with a Normalization and Dropout layers followed by an individual Dense layer with SoftMax activation function and 15 outputs representing the 15 classes.In the Figure 3, CNN architecture of our model is represented and some acronyms used in the figure are C - conv2d, R - Relu, B - batch normalizer, MP - max pooling, DO - drop out Visualization: To understand how the model works and what is exactly being learned, the intermediate layer activations that display the various feature maps output by various Convolution, Normalization, and Max Pooling layers in a network when a certain input is given are visualized. This helps to get an idea of how the input image is decomposed into different filters by the network. Figure 4,5,6 shows the various layers outputs.

Figure 3: CNN architecture

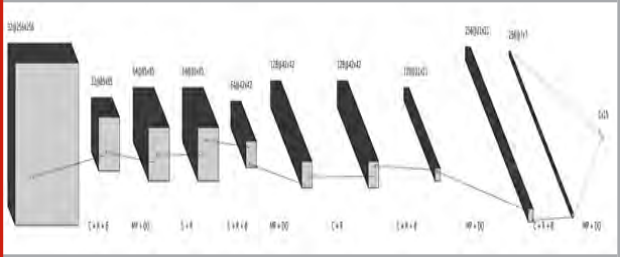


Table 1. Model Summary

Layer (type)	Output Shape	Param #
conv2d_1 (Conv2D)	(None, 256, 256, 32)	896
activation_1 (Activation)	(None, 256, 256, 32)	0
batch_normalization_1 (Batch	(None, 256, 256, 32)	128
max_pooling2d_1 (MaxPooling2	(None, 85, 85, 32)	0
dropout_1 (Dropout)	(None, 85, 85, 32)	0
conv2d_2 (Conv2D)	(None, 85, 85, 64)	18496
activation_2 (Activation)	(None, 85, 85, 64)	0
conv2d_3 (Conv2D)	(None, 85, 85, 64)	36928
activation_3 (Activation)	(None, 85, 85, 64)	0
batch_normalization_2 (Batch	(None, 85, 85, 64)	256
max_pooling2d_2 (MaxPooling2	(None, 42, 42, 64)	0
dropout_2 (Dropout)	(None, 42, 42, 64)	0
conv2d_4 (Conv2D)	(None, 42, 42, 128)	73856
activation_4 (Activation)	(None, 42, 42, 128)	0
conv2d_5 (Conv2D)	(None, 42, 42, 128)	147584
activation_5 (Activation)	(None, 42, 42, 128)	0
batch_normalization_3 (Batch	(None, 42, 42, 128)	512
max_pooling2d_3 (MaxPooling2	(None, 21, 21, 128)	0
dropout_3 (Dropout)	(None, 21, 21, 128)	0
conv2d_6 (Conv2D)	(None, 21, 21, 256)	295168
activation_6 (Activation)	(None, 21, 21, 256)	0
batch_normalization_4 (Batch	(None, 21, 21, 256)	1024
max_pooling2d_4 (MaxPooling2	(None, 7, 7, 256)	0
dropout_4 (Dropout)	(None, 7, 7, 256)	0
flatten_1 (Flatten)	(None, 12544)	0
dense_1 (Dense)	(None, 1024)	12846080
activation_7 (Activation)	(None, 1024)	0
batch_normalization_5 (Batch	(None, 1024)	4096
dropout_5 (Dropout)	(None, 1024)	0
dense_2 (Dense)	(None, 15)	15375
activation_8 (Activation)	(None, 15)	0

Figure 4: Full-Color intermediate activation in the Convolution layer

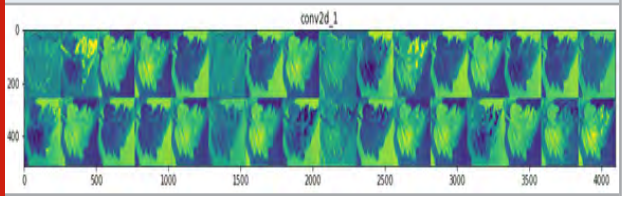


Figure 5: Full-Color intermediate activation in Pooling layer

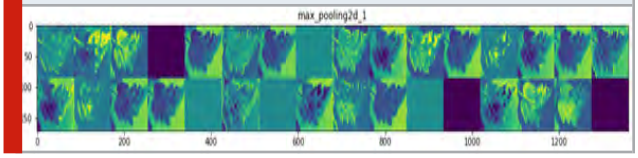


Figure 6: Full-Color intermediate activation in Normalization layer

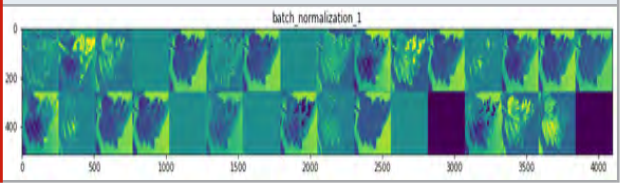


Table 2. Experimental Result

Activation Function	Optimizer	Learning Rate	# of Epochs	Batch Size	Loss	Accuracy in %
Relu	adadelta(rho=0.95)	1.0	20	64	1.3584	62.70
Relu	adadelta(rho=0.95)	1.0	10	32	0.0446	89.32
Relu	adadelta(rho=0.95)	1.0	10	32	0.3113	88.43
Relu	adadelta(rho=0.95)	1.0	10	32	0.1344	87.17
Relu	adadelta(rho=0.95)	1.0	9	32	0.0054	88.05
Relu	adadelta(rho=0.95)	1.0	30	32	0.4531	87.12
Relu	adamax	0.002	20	64	1.8768	62.10
Relu	Adam	0.001	20	64	0.5292	84.54
Relu	Nadam	0.002 beta1=.9 beta2=.99	20	64	0.9715	74.02
Relu	Nadam	0.002 beta1=.9 beta2=.99	20	32	1.2239	71.76
Relu	Nadam	0.002 beta1=.9 beta2=.99	10	32	0.7229	79.92

RESULTS

Different CNN models with different attributes were trained and the model in [Table 1] yielded the better results. Once a proper model was obtained, learning parameters of that model was varied and different accuracies as tabulated in the Table 2. That model under the setting of Optimizer: adadelta (1.0, rho=0.95), Loss function: Categorical Cross Entropy, train dataset ratio of 70% and test dataset ratio of 30% yielded the best accuracy of 89.325% on full color images.

CONCLUSION

In this article, CNN based leaf disease detection method is proposed. These neural network models can be trained to detect diseases across a largely diverse species of plants and can grow proportional to the size of the dataset provided. In addition these models can be integrated in a mobile application and can be accessed without internetwork connectivity and the mobile application also displays suggestions of what pesticide to use, thus making it possible to use this system with very less computing power and in areas with poor or no network connections and hence enabling the use of technology in areas of agriculture and other flora based ecosystems on the go without much constraints.

REFERENCES

Ashqar, Belal and Abu-Naser, Samy. (2018). Image-Based Tomato Leaves Diseases Detection Using Deep Learning. International Journal of Academic Engineering Research. 2(12). 10-16

Ghulam Mustafa Choudhary and Vikrant Gulati (2015). Advance in Image Processing for Detection of Plant

Diseases. International Journal of Advanced Research in Computer Science and Software Engineering 5(7), 1090-1093

Hrishikesh P. Kanjalkar, and S.S.Lokhande (2014). Feature Extraction of Leaf Diseases International Journal of Advanced Research in Computer Engineering & Technology. 3(1). 153-155

Image Processing and Support Vector Machine(SVM), Journal for Research, 02(02), 74 – 77.

Plant Leaf Disease Detection and Classification using Image Processing Techniques. International Journal of Innovative and Emerging Research in Engineering. 2(4) ,. 139-144.

PradnyaRavindra Narvekar, Mahesh Manik Kumbhar, & S N Patil (2014). Grape Leaf Diseases Detection & Analysis using SGDM Matrix Method. International Journal of Innovative Research in Computer and Communication Engineering. 2(3). 3365-3372.

Prajakta Mitkal, Priyanka Pawar, Mira Nagane, Priyanka Bhosale, Mira Padwal and Priti Nagane (2016). Leaf Disease Detection and Prevention Using Image processing using MATLAB. 2(2), 26-30.

Prakash M. Mainkar, Shreekant Ghorpade, and Mayur Adawadkar (2015).

Sabah Bashir and Navdeep Sharma (2012). Remote Area Plant Disease Detection using Image Processing. IOSR Journal of Electronics and Communication Engineering (IOSRJECE). 2(6), 31-34

Sen Maitra, Durjoy & Bhattacharya, Ujjwal and Parui, Swapan. (2015). CNN based common approach to handwritten character recognition of multiple scripts. 1021-1025.

Sharada Prasanna Mohanty (2016). Using Deep Learning for Image-Based Plant Disease Detection. Frontiers in Plant Science, 1-10.

Sujatha R, Sravan Kumar Y and Garine Uma Akhil (2017). Leaf disease detection using image processing. Journal of Chemical and Pharmaceutical Sciences. 10(1), 670-672.

Vaijinath B Batule, Gaurav U Chavan, Vishal P Sanap and Kiran D Wadkar (2016), Leaf Disease Detection using

Vijai Singh,, A. K. Mishra (2017). Detection of plant leaf diseases using image segmentation and soft computing techniques. Information Processing In Agriculture, 4, 41–49.

Biosc.Biotech.Res.Comm. Special Issue Vol 13 No 11 (2020) Pp-88-93

BBRC
Bioscience Biotechnology Research Communications

Plant Disease Detection System for Smart Agriculture

R. Indhu[1] and K.Thilagavathi[2]

[1]*PG student, Department of ECE, Kumaraguru College of Technology, Coimbatore, India*

[2]*Asst Professor, Department of ECE, Kumaraguru College of Technology, Coimbatore, India*

ABSTRACT

Indian economy relies on agriculture to a greater extent. In traditional agriculture, the farmers identify the crop diseases with the help of an expert either by direct visual inspection or by sending the diseased images to experts through online services. Also, continuous monitoring cannot be done manually. The main objective is to develop an android application which identifies and classifies three major diseases - Black horse riding, Brown spot and Bacterial leaf steak. In the plant disease detection system, image to be tested is acquired, pre-processed, segmented and classified based on the disease type. The classification is performed using probabilistic linear classifier called Naive Bayes. The application is developed using Android Studio and the programming language used for the development is java. This application identifies the plant disease based on pixel intensities, predicts the plant growth and sunlight condition if it is good or not. It suggests suitable fertilizers and pesticides depending on the disease type. The average accuracy of the developed application is about 80% and the implementation of this system reduces manpower and increases productivity.

KEY WORDS: ANDROID APPLICATION, IMAGE PROCESSING, NAIVE BAYES CLASSIFIER, PLANT DISEASES.

INTRODUCTION

Human survival is greatly influenced by the food supply from nature, which purely depends upon agriculture. Since the population increases day-by-day, the demand for food also increases. Hence, agriculture is a field that never ends at all. Crop diseases leads to less productivity which in turn affects the crop yield. Crop yield can be increased by detecting the diseases in the early stages. The traditional method of disease detection and classification involves visual inspection, which is expensive, time consuming and sometimes provides incorrect results. Hence digital image processing techniques and probabilistic neural network algorithm-Naive Bayes are applied to improve the accuracy of detecting and classifying the crop diseases in an unmanned way. The main objective is to develop an android application which acquires image, identifies the disease type, and suggests the fertilizers and pesticides. This increases the crop yield and reduces manpower and labor cost. In India, rice and wheat are the most widely used food crops. The demand for rice and wheat increases day-by-day. The main obstacle for achieving the target is the bacterial and fungal diseases. In severe cases, these diseases can lead to yield loss up to 50%. This paper considers three major bacterial and fungal diseases affecting mostly rice and wheat. These diseases are also present in maize, sugarcane, palm, ornamental plants etc.

The two diseases of the rice plant; Leaf Blast and Brown Spot are classified using Self Organizing Map (SOM), an unsupervised learning technique with an accuracy of approximately 92% (Phadikar and Sil,2008). The authors used multiclass Support Vector Machine (SVM) with Gaussian kernel function for classifying three types of diseases in paddy (Shah, Prajapati and Dabhi, 2016). Fuzzy classifier is used for classifying four different diseases in wheat plant. The accuracy for disease detection and disease type classification are 88% and 56% respectively (Diptesh et al., 2014). Artificial

ARTICLE INFORMATION

*Corresponding Author: indhu.18mco@kct.ac.in
Received 15th Oct 2020 Accepted after revision 5th Dec 2020
Print ISSN: 0974-6455 Online ISSN: 2321-4007 CODEN: BBRCBA

Thomson Reuters ISI Web of Science Clarivate Analytics USA and Crossref Indexed Journal

NAAS Journal Score 2020 (4.31) SJIF: 2020 (7.728)

Online Contents Available at: http//www.bbrc.in/
Doi: http://dx.doi.org/10.21786/bbrc/13.11/20

Neural Network (ANN) and SVM classifiers are used for detection and classification of maize diseases using python. ANN has resulted an accuracy of 55% to 65% and SVM has resulted an accuracy of 70% to 75% (Durga and Anuradha, 2019).Fuzzy expert system is designed and developed for identifying nine different diseases of finger millets (Roseline, Tauro and Ganesan, 2012). The system identifies the diseases by analysing the symptoms for each disease and provides tips to overcome the diseases.

An approach for identifying a pearl millet disease was introduced using Convolutional Neural Network (CNN) model VGG16 and obtained an accuracy of 95% (Coulibaly et al., 2019). Alexnet is used for identifying 14 different diseases of olive plant for which 80% of the dataset for training and the remaining 20% for testing is utilized and obtained an accuracy of 99.11% (Alruwaili et al., 2019). A system for detecting diseases in palm leaves is developed using multiclass SVM classifier. The system detects two diseases: Anthracnose and Chimaera with an accuracy of 96% (Masazhar and Kamal, 2018). A system is proposed where four different diseases of groundnut leaves are detected and classified using Back propagation network in MATLAB and obtained an accuracy of 97.41% (Ramakrishnan and Sahaya, 2015).

Algorithms like Multiclass SVM, Naïve Bayes, K-Nearest Neighbour and Multinomial Logistic Regression were used to classify three different diseases of sunflower with an accuracy of 92.15%, 89.075%, 89.32% and 92.57% respectively. The accuracy of classifying the healthy leaves are 100% in all the four classification algorithms (Pinto et al., 2016). Healthy leaves and scorch leaves are classified using K-Nearest Neighbour algorithm with K=1 in MATLAB and obtained an accuracy of about 95% (Eaganathan et al., 2014). An automated system device is designed to detect Mildew disease from healthy plants and to estimate the severity of the disease. K-Nearest Neighbour algorithm is used for detecting the disease and used decision tree for severity estimation. The accuracy of the experiment is 82.5% (Parikh et al., 2016). Based on the different parts of plant, 18 classes of diseases were trained and tested using Deep Convolutional Neural Network (DCNN).

Six models were designed and out of which, the performance of ResNet50 and InceptionV2 based models were better. The accuracy of the developed system is about 90% (Selvaraj et al., 2019). An image processing system is developed using SVM to identify two major diseases of tea namely Algal and Brown Blight from the healthy tea leaves (Hossain et al., 2018). The success rate of the proposed system was about 93% in which 50 black coffee beans and 50 healthy coffee beans were used. They used grey scale image for colour feature extraction and K-means clustering for segmentation process. The results obtained are 100% accurate using MATLAB software (Shetty, Kotian and Uppar, 2019). An application which classifies two diseases- coffee leaf miner and coffee leaf rust from the normal coffee leaves is developed using Back propagation algorithm. The application also estimates the severity of disease (Manso et al., 2019). Texture Based Disease Recognition experiment is conducted and calculated kappa coefficient and sensitivity as 0.900 and 0.933, respectively. Similarly, they calculated the same parameters using Deep Learning Disease Recognition and obtained the result as 0.970 and 0.980 respectively (Boa Sorte et al., 2019).

The system classifies two cucumber crop diseases namely powdery mildew, downy mildew from the healthy using Artificial Neural Network with the classification accuracy of 80.4% (Pooja Pawar et al, 2016).Graphical User Interface (GUI) is also developed to indicate the diagnosis and treatment for the detected disease (Pawar, Turkar and Patil, 2016). Backpropagation neural network is used for classifying healthy pomegranate leaves from the diseased leaves and obtained an accuracy of 90% (Pawar and Jadhav, 2018). The leaves of 3 commonly attacked disease of grape are experimented using MATLAB. The feature extraction using LAB color model given an accuracy of 82.5% and the feature extraction using both LAB and HSI color model gives an accuracy of 90%. Multiclass SVM is used for classification of diseases (Agrawal, Singhai and Agarwal, 2018). Classification of two grape diseases downy and powderly was done using SVM classifier with 88.89% accuracy (Padol and Yadav, 2016). This paper focuses on three main diseases that affects different plants namely - Bacterial leaf streak, Brown spot and Black horse riding.

Bacterial Leaf Streak: Bacterial leaf streak is one of the bacterial diseases which attacks mostly wheat plant. Symptoms initially appear as water-soaked streaks and turns into brown linear lesions between leaf veins. Initially, these leaf streaks are dark green in color and when the disease becomes severe, it turns to brownish or yellowish gray. The complete leaves may become brown and dead when the disease is extremely severe. Applying copper-based fungicide on leaves or pesticide named Epsom salt can help in controlling the disease. The yield loss due to bacterial leaf streak disease is about 10% to 40%. Bacterial leaf streak disease is shown in figure 1.

Figure 1: Bacterial leaf streak disease

Brown Spot: Brown spot, a fungal disease occurs mostly in rice and soya bean plants. Infected seedlings have

brown lesion, and the leaves turn to rusty brown. The diseased seedlings die at the early stage or become underdeveloped. Fungicides like triadimefon can help in the reduction of this disease. Leaf wetness can increase the severity of the disease within 6 to 36 hours. The yield loss due to brown spot disease is about 8% to 15% but premature defoliation of canopy may increase the yield loss of 25% to 50%. Brown spot disease is shown in figure 2.

Figure 2: Brown spot disease

Black Horse Riding: Black horse-riding disease mostly attacks the monocotyledons. Most of the patches caused by this disease are due to pathogenic fungi. Once the fungi enter the leaf, it continues to grow and destroys the leaf tissue. This results in spots of varying sizes. The dead areas on the leaves will be dark in color. The border of the malignant areas in the leaf will be in red or purple. When the disease becomes severe, it leads to complete defoliation of leaves. A chemical fungicide or any number of organic options such as Copper, Lime Sulphur, Neem Oil, Potassium or Ammonium Bicarbonate can be used as fertilizer. Black horse-riding disease is shown in figure 3.

Figure 3: Black horse riding disease

MATERIAL AND METHODS

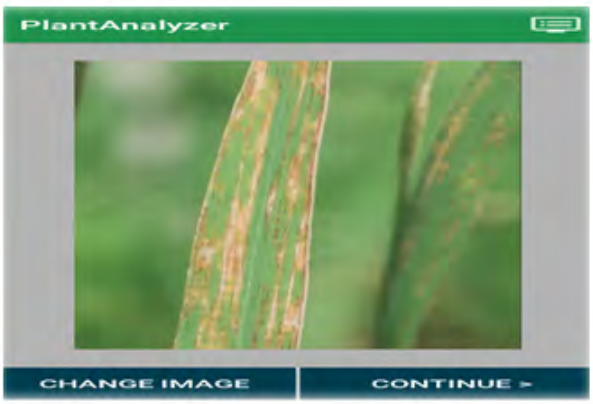
An android application is developed using Android Studio through which the images are acquired, pre-processed, segmented and classified based on the disease.

Figure 4 represents the flow of the proposed system for plant disease classification.

Figure 4: Proposed system for plant disease classification

Image Acquisition: In image acquisition, the colored image of the diseased or healthy leaf to be experimented is acquired through high clarity mobile phone or from the internet. The leaves with different canopy size were selected to decrease the difficulty of the classification problem. In the developed Android application, image to be tested is chosen from the gallery as in figure 5.

Figure 5: Selected Image for testing



Preprocessing: Image pre-processing is a technique wherein the unwanted noises present in the images are removed. In the proposed work, there are three steps in image pre-processing - RGB to Gray color transformation, Bitmap extraction and Histogram analysis.

RGB to Gray color transformation: The first step in pre-processing is to convert the colored image of leaf into gray color. RGB images will have three different intensities for each pixel. So, on converting it to gray color, there will be only single intensity for each pixel.

Bitmap Extraction: The second step in pre-processing is the extraction of bitmap so that the required part of the leaf is separated from the background.

Histogram Analysis: The image thus obtained is analyzed with the pixel intensity in order to form the histogram graph. In the histogram graph, the left most area of the horizontal axis denotes the dark tone of the image and similarly the right side denotes the light tone of the image.

Segmentation: Image segmentation is the process of fragmenting an image into several segments. This is done to make the analysis easier. After the histogram analysis, threshold-based segmentation is done. The output of this step will be a binary image. This is the simplest method used for image segmentation, but it is still effective. The segmented image is shown in figure 6.

Figure 6: Image segmentation

Disease detection and classification: After the segmentation process, images are classified based on Naive Bayes algorithm. This algorithm obeys Bayes theorem. It comes under the category of probabilistic classifiers. It needs only a smaller number of data sets compared to other classifiers. There will be different pixel values for each disease. The input image to be tested compares with all possible diseases after the segmentation process. It predicts the probability for all the possible diseases and the output will be the disease which has the highest probability. The application also displays the percentage of the diseased portion.

Growth prediction: Based on the size and shape of the leaf, its growth is predicted. The application also predicts the sunlight condition if it is good or not. If the percentage of disease is more than 5%, then the sunlight condition is not good. Based on the disease type, the android application suggests the fertilizers and pesticides.

Figure 7: Input image to be tested

RESULTS AND DISCUSSION

The android application developed is installed in the android device. The image to be tested is selected from the gallery. The app has two options – one is to change the image and the other is to continue. These two choices are fixed using the Android studio press. The application screen of the same is shown in figure 7.

Once continue is pressed, it will move to the next page of the application. In this step, the given input image checks with all three diseases Bacterial leaf streak, brown spot, and black horse riding. The portion matching with the diseases will be red in color as in figure 8. If report option is pressed, then the disease that has the maximum probability will be displayed along with the percentage of infection as shown in figure 9. A button is inserted in the bottom of this page. Once this button is pressed, the application will predict the sunlight condition if it is good or not and predicts the growth of the plant. It also suggests the fertilizers and pesticides based on the type of the disease as shown in figure 10. A button is inserted in that page for sharing these suggestions.

For each disease, five images were tested. For bacterial leaf streak, all the five images were classified correctly and its accuracy is 100%. In case of brown spot, three images were identified correctly out of five images with the accuracy 60% and for Black horse riding disease, four images were identified correctly with the accuracy of 80%.The average accuracy obtained is 80%.The graphical

representation of the classified image accuracy is shown in Figure 11. Here, the x-axis denotes the disease type and the y-axis denotes the accuracy.

Figure 8: Possible diseases

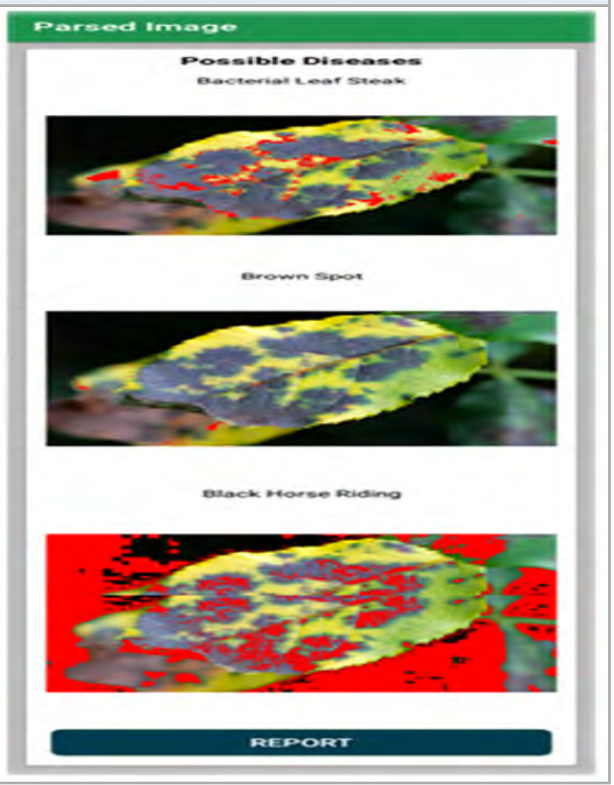


Figure 10: Fertilizers and pesticides suggestion

Figure 9: Disease type

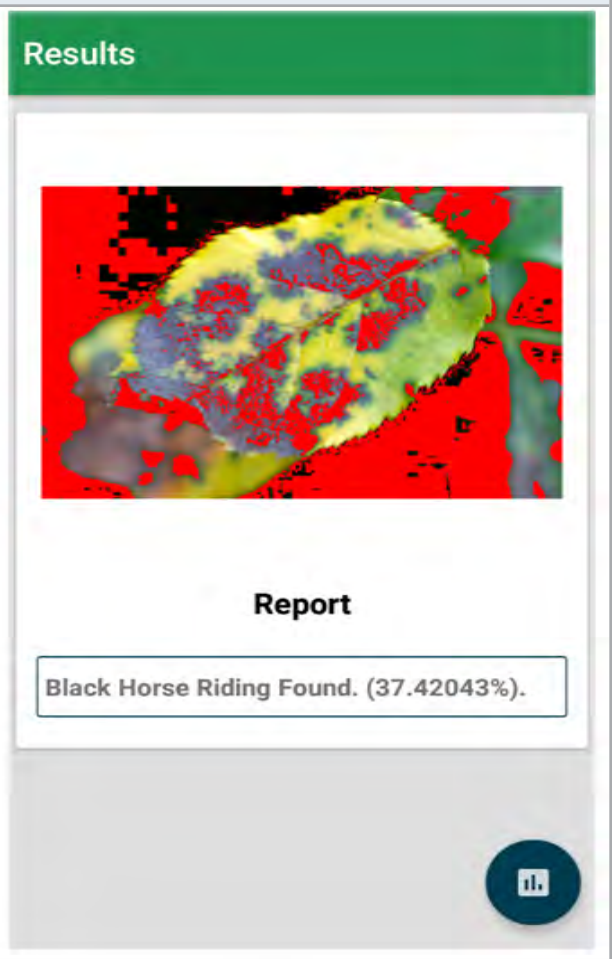
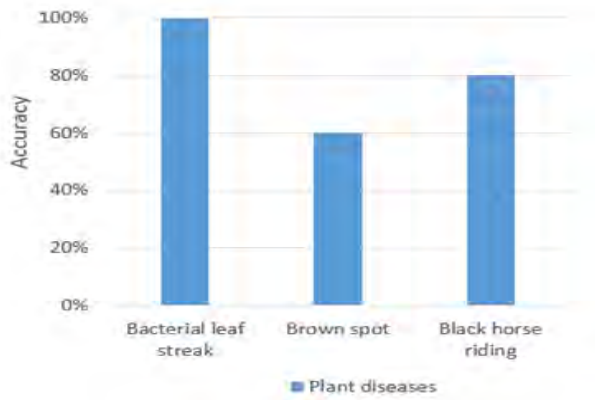


Figure 11: Performance Analysis



CONCLUSION

The developed android application uses RGB to gray color transformation, bitmap extraction and histogram analysis for image pre-processing. Thresholding based segmentation is done and finally disease type is classified using Naïve Bayes classifier. In this experiment, three class of plant diseases – Bacterial leaf streak, Brown spot and Black horse riding are considered. For testing

purpose, 15 images are used. The average accuracy of the application is 80%. In future, other algorithms like decision tree can be applied to improve the accuracy.

ACKNOWLEDGEMENTS

Nil

Conflict of interest: The author declares that there is no conflict of interest.

REFERENCES

Agrawal, N., Singhai, J. and Agarwal, D. K. (2018) Grape leaf disease detection and classification using multi-class support vector machine, International Conference on Recent Innovations in Signal Processing and Embedded Systems, RISE 2017. IEEE, 2018-Janua, pp. 238–244. doi: 10.1109/RISE.2017.8378160.

Alruwaili, M. et al. (2019) An efficient deep learning model for olive diseases detection, International Journal of Advanced Computer Science and Applications, 10(8), pp. 486–492. doi: 10.14569/ijacsa.2019.0100863.

Boa Sorte, L. X. et al. (2019) Coffee Leaf Disease Recognition Based on Deep Learning and Texture Attributes, Procedia Computer Science, 159, pp. 135–144. doi: 10.1016/j.procs.2019.09.168.

Coulibaly, S. et al. (2019) Deep neural networks with transfer learning in millet crop images, Computers in Industry, 108, pp. 115–120. doi: 10.1016/j.compind.2019.02.003.

Diptesh, M. et al. (2014) Application of Fuzzy C-Means Clustering Method to Classify Wheat Leaf Images based on the presence of rust disease, Advances in Intelligent Systems and Computing, 327(October). doi: 10.1007/978-3-319-11933-5.

Durga, N. K. and Anuradha, G. (2019) Plant disease identification using SVM and ANN algorithms, International Journal of Recent Technology and Engineering, 7(5), pp. 471–473.

Eaganathan, U. et al. (2014) Identification of Sugarcane Leaf Scorch Diseases using K-means Clustering Segmentation and K-NN based Classification, International Journal of Advances in Computer Science and Technology, 3(12), pp. 11–16.

Hossain, M. S. et al. (2018) Recognition and detection of tea leaf's diseases using support vector machine, Proceedings - 2018 IEEE 14th International Colloquium on Signal Processing and its Application, CSPA 2018, (March), pp. 150–154. doi: 10.1109/CSPA.2018.8368703.

Manso, G. L. et al. (2019) A smartphone application to detection and classification of coffee leaf miner and coffee leaf rust. Available at: http://arxiv.org/abs/1904.00742.

Masazhar, A. N. I. and Kamal, M. M. (2018) Digital image processing technique for palm oil leaf disease detection using multiclass SVM classifier, 2017 IEEE International Conference on Smart Instrumentation, Measurement and Applications, ICSIMA 2017, 2017-November, pp. 1–6. doi: 10.1109/ICSIMA.2017.8311978.

Padol, P. B. and Yadav, A. A. (2016) SVM classifier based grape leaf disease detection, Conference on Advances in Signal Processing, CASP 2016, pp. 175–179. doi: 10.1109/CASP.2016.7746160.

Parikh, A. et al. (2016) Disease detection and severity estimation in cotton plant from unconstrained images, Proceedings - 3rd IEEE International Conference on Data Science and Advanced Analytics, DSAA 2016, pp. 594–601. doi: 10.1109/DSAA.2016.81.

Pawar, P., Turkar, V. and Patil, P. (2016) Cucumber disease detection using artificial neural network, Proceedings of the International Conference on Inventive Computation Technologies, ICICT 2016, 2016. doi: 10.1109/INVENTIVE.2016.7830151.

Pawar, R. and Jadhav, A. (2018) Pomogranite disease detection and classification, IEEE International Conference on Power, Control, Signals and Instrumentation Engineering, ICPCSI 2017. IEEE, pp. 2475–2479. doi: 10.1109/ICPCSI.2017.8392162.

Phadikar, S. and Sil, J. (2008) Rice disease identification using pattern recognition techniques, Proceedings of 11th International Conference on Computer and Information Technology, pp. 420–423. doi: 10.1109/ICCITECHN.2008.4803079.

Pinto, L. S. et al. (2016) Crop Disease Classification using Texture Analysis,IEEE International Conference on Recent Trends in Electronics Information Communication Technology, pp. 825–828.

Ramakrishnan, M. and Sahaya, A. N. A. (2015) Groundnut leaf disease detection and classification by using back probagation algorithm, 2015 International Conference on Communication and Signal Processing, ICCSP 2015, (7092512506), pp. 964–968. doi: 10.1109/ICCSP.2015.7322641.

Roseline, P., J.M Tauro, C. and Ganesan, N. (2012) Design and Development of Fuzzy Expert System for Integrated Disease Management in Finger Millets, International Journal of Computer Applications, 56(1), pp. 31–36. doi: 10.5120/8857-2815.

Selvaraj, M. G. et al. (2019) AI-powered banana diseases and pest detection, Plant Methods. BioMed Central, 15(1), pp. 1–11. doi: 10.1186/s13007-019-0475-z.

Shah, J. P., Prajapati, H. B. and Dabhi, V. K. (2016) A survey on detection and classification of rice plant diseases, IEEE International Conference on Current Trends in Advanced Computing (February 2019). doi: 10.1109/ICCTAC.2016.7567333.

Shetty, B., Kotian, K. R. and Uppar, A. S. (2019) Coffee Bean Detection and Segregation, International Journal of Engineering Science and Computing, 9(4), pp. 21636–21638.

Biosc.Biotech.Res.Comm. Special Issue Vol 13 No 11 (2020) Pp-94-98

BBRC
Bioscience Biotechnology Research Communications

Prediction of Autism Spectrum Disorder Using Rough Set Theory

V.Geetha[1] and V. Jalaja Jayalakshmi[1]
[1]*Department of Computer Applications, Kumaraguru College of Technology, Coimbatore, India*

ABSTRACT

Autism Spectrum Disorder (ASD) is a neurological disease that starts early in childhood and persists throughout a person's life. It is a condition linked with brain development and influences a person's behaviour and their interaction with others. Autism has a wide range of symptoms which can vary from person to person. There is no direct medical test to diagnose ASD disorder and hence trained physicians are needed to oversee the person's behaviour development to detect it. There is no cure for ASD, and early detection of the illness will be able to make significant quality improvements in the behaviour of the affected person. Machine Learning techniques are widely used to identify the factors associated with the disease, thus helping in early detection. This paper attempts to explore the possibilities of analyzing the autism data sets of adults using rough set theory and predict the main factors associated with the disorder for providing an early treatment. A comparative performance analysis of the results is done using two rough set algorithms, and the results indicate that the genetic algorithm gives a better performance in this domain.

KEY WORDS: AUTISM SPECTRUM DISORDER, MACHINE LEARNING, RECEIVER OPERATING CHARACTERISTICS (ROC) CURVE, REDUCTS, ROUGH SETS.

INTRODUCTION

Autism is a brain developmental disorder that starts early in childhood and affects the person's behavior amongst several age groups. It has a wide range of symptoms that varies from person to person. The prevalence of autism is increasing at an alarming rate and has mainly been found to occur in boys more than girls. Research shows that currently, autism in boys occurs at a 4:1 ratio and in girls, a 7:1 ratio. The overall increase in worldwide autism reporting can be attributed to a general rise in awareness of the disorder, improving technology that can diagnose the disease, and overall detectability measures that have allowed healthcare professionals to identify the condition at an earlier age. However, it is also plausible to say that unknown risk factors that occur in day-to-day life contribute to the rise in autism cases around the world and, to prove this point, further research is needed to elaborate on this stance. Data mining is the method of finding patterns from vast amounts of data stored in large repositories. (Han and Kamber, 2001). These methods can handle and process large-scale information (Reidmiller,1994). The modeling techniques of data mining include classification, clustering, regression, sequential patterns etc.

Classification assigns objects to a predefined class and is an essential task of data mining. A single classification algorithm will not be able to provide good results in all cases. The choice of an optimal classification algorithm highly depends on the problem domain. Data mining has great potential in the health care industry and is becoming more and more popular. Medical databases store a lot of information about the patients and their

ARTICLE INFORMATION

*Corresponding Author: geetha.v.mca@kct.ac.in
Received 15th Oct 2020 Accepted after revision 4th Dec 2020
Print ISSN: 0974-6455 Online ISSN: 2321-4007 CODEN: BBRCBA

Thomson Reuters ISI Web of Science Clarivate Analytics USA and Crossref Indexed Journal

NAAS Journal Score 2020 (4.31) SJIF: 2020 (7.728)

Online Contents Available at: http//www.bbrc.in/
Doi: http://dx.doi.org/10.21786/bbrc/13.11/21

medical conditions. Classification of disease disorders from these medical databases is very challenging with standard data mining algorithms because knowledge is incomplete (Hassanien and Ali, 2004). Various algorithms are used to classify the autism data, and the authors have concluded that Multilayer perceptron algorithm gives the highest classification accuracy (Jalaja et al,. 2019).

The use of Rough set methodology in the data mining field has increased. Rough set theory was initially proposed to address vagueness in classifying the objects in each set (Pawlak, 1998). It is based on the principle that every object has some information associated with it, and objects having same information are similar and can be categorized to the same class. Rough set applications are widely used today in the medical, educational fields and other industrial applications. Rough sets are applied for bioinformatics related data (Cao et al., 2006) and to biomedical datasets (Revett, 2005) and have generated a highly accurate classifier. A rule-based disease identification system from ECG has been created (Mitra et al. 2006). The behavior of decision trees and rough set algorithms have been examined (Perzyk and Soroczynski, 2010) in industrial applications, and it has been found out that rough set algorithms can be used when a detailed set of rules are required. Rough sets have been compared with decision trees in several applications (Daubie et al., 2002, Mak and Munakata, 2002). Rough set algorithms are used to classify students' performance in adaptive E-assessment (Nandakumar et al., 2014).

Since autism has no known cure, behavioral studies are the primary way to detect autism currently; further research on these lines would prove to be beneficial to detect and possibly slow down the progression of this developmental disorder. The aim of this paper is to use rough set theory on the UCI Autism adult data set for forecasting autistic symptoms at an early stage. The organization of the paper is as follows: the theoretical aspects of the rough set algorithms are discussed in Section 2. Section 3 discusses experimental setup, results, and analysis. Future work is discussed in Section 4.

MATERIAL AND METHODS

Rough Set Theory: A Rough set is a mathematical method to analyze data stored in tabular form (Dubois and Prade 1991). Rough Set Theory finds equivalence classes from the training data. The objects that form the equivalence class are indiscernible, i.e. are identical about the attributes that explains the data.

Johnson's Algorithm: Johnson's algorithm follows a greedy strategy using heuristics for generating the reducts. It begins with an empty set S and counts the occurrence of each of the attributes within the discernibility clause in each iteration. The attribute that has the maximum count 'f' is added into S and is removed from the clause. The algorithm continues till all the attributes in the clause are removed from the discernibility function, and it returns S, which contains the reduct. This algorithm will generate a single reduct with a minimum number of attributes and does not support approximate solutions.

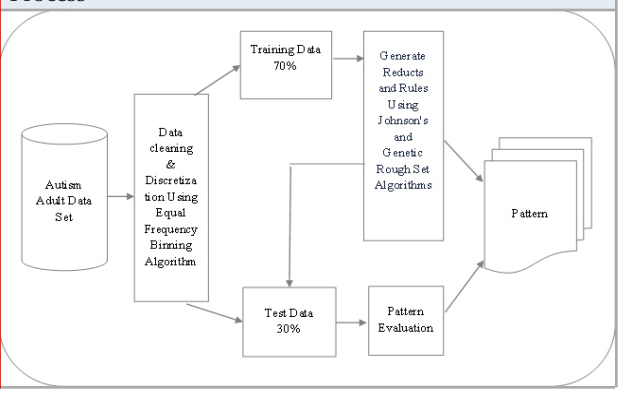
Genetic algorithm: Genetic Algorithm (GA) is based on heuristic approach and is inspired by organism evolution. It generates a large set of potential solutions to a problem. It evaluates the obtained solution using a fitness function. These results produce new solutions. The parent solutions with improved fitness solutions have better chances of reproduction. The reducts generated using GA begins with initializing the chromosomes from the rough set discernibility clause. The three basic processes of the genetic algorithm are generating an initial population, fitness function evaluation and reproduction.

RESULTS AND DISCUSSION

Experimental Setup: The proposed work was carried out using an open-source data set," Autism Adult data set" from the UCI Machine Learning Repository (Dua D and Graff C, 2019), which had 704 observations with 21 variables. The data set was analyzed using rough set theory. The results obtained using different algorithms in rough set theory were compared using various prediction measures. Rosetta is a toolkit application using the rough set methodology used for analyzing data stored in the form of tables.

A framework for rough set-based classification process: A rough set-based model has been developed on autism adult data set for finding decision rules to predict the occurrence of Autism Spectrum Disorder using Rosetta rough set tool. The steps in the application of rough set process are shown in Figure 1.

Figure 1: A Framework For Rough Set-Based Classification Process



Data Preprocessing: The data set has 21 variables, out of which only two variables (age and result) are numeric, and the remaining variables are categorical/binary. There are 704 observations, out of which 189 has been classified as 'having ASD' and the remaining 515 as 'not having ASD'. It has been observed that there is a bias towards the number of observations not having ASD. The attributes 'age-desc' and 'used-app before' has the same value for all the observations and will not contribute much towards decision making. Hence, they

are removed from the autism data set using the Rosetta tool. The data set has 15 attributes and 1 class variable after the cleaning step.

In data discretization, each continuous variable is converted into an interval-based value. Rosetta tool has a few discretization algorithms, the numerical attribute age is discretized using equal frequency binning algorithm and is shown in Figure 2. After discretization, the database is arbitrarily divided into two different datasets namely training and test data sets. A model is built using the training set and is then evaluated on the test set. Different sample sizes are randomly selected from the data set for various split factors (seed) ranging from 0.1 to 0.9 and are used for training and testing purposes.

Figure 2: Data Discretization By Equal Frequency Binning Algorithm

age	gender	ethnicity	Jaundice	contry_of_res	ASD
[24, 32)	f	White-Europ	no	'United State	NO
[24, 32)	m	Latino	no	Brazil	NO
[24, 32)	m	Latino	yes	Spain	YES
[32, *)	f	White-Europ	no	'United State	NO
[32, *)	m	Others	yes	'United State	YES
[*, 24)	f	Black	no	'United State	NO
[32, *)	m	White-Europ	no	'New Zealan	NO
[24, 32)	m	White-Europ	no	'United State	NO
[*, 24)	m	Asian	yes	Bahamas	YES
[32, *)	m	White-Europ	no	'United State	YES
[*, 24)	f	'Middle Easte	no	Burundi	NO
[*, 24)	m	'Middle Easte	no	'New Zealan	NO
[24, 32)	m	'Middle Easte	no	Jordan	NO
[24, 32)	m	White-Europ	no	Ireland	NO
[32, *)	f	'Middle Easte	no	'United Arab	NO

Figure 3: Few Reducts Generated Using Johnson's Algorithm

Johnson 40

	Reduct	Support	Length
1	{A2_Score, A3_Score, A5_Score, A6_Score, A7_Score, A9_Score, ethnicity, contry_of_res}	100	8

Figure 4: Few Reducts Generated Using Genetic Algorithm

	Reduct	Support	Length
2	{A2_Score, A3_Score, A4_Score, A5_Score, A6_Score, A7_Score, A9_Score, ethnicity, contry_of_res}	100	9
3	{A1_Score, A4_Score, A5_Score, A6_Score, A8_Score, A10_Score, gender, ethnicity, contry_of_res}	100	9
4	{A1_Score, A4_Score, A5_Score, A6_Score, A7_Score, age, gender, ethnicity, contry_of_res}	100	9
5	{A1_Score, A3_Score, A5_Score, A6_Score, A9_Score, A10_Score, gender, ethnicity, contry_of_res}	100	9
6	{A2_Score, A5_Score, A6_Score, A7_Score, A8_Score, A9_Score, gender, ethnicity, Jaundice, contry_of_res}	100	10
7	{A1_Score, A3_Score, A4_Score, A6_Score, A7_Score, A9_Score, age, ethnicity, Jaundice, contry_of_res}	100	10
8	{A1_Score, A2_Score, A3_Score, A4_Score, A6_Score, A9_Score, A10_Score, age, ethnicity, contry_of_res}	100	10
9	{A1_Score, A4_Score, A5_Score, A6_Score, A8_Score, A9_Score, age, gender, ethnicity, contry_of_res}	100	10
10	{A1_Score, A3_Score, A6_Score, A7_Score, A8_Score, A10_Score, age, gender, ethnicity, contry_of_res}	100	10
11	{A1_Score, A2_Score, A3_Score, A4_Score, A5_Score, A6_Score, A8_Score, A9_Score, ethnicity, contry_of_res}	100	10
12	{A2_Score, A3_Score, A4_Score, A6_Score, A7_Score, A8_Score, A9_Score, A10_Score, age, contry_of_res}	100	10
13	{A1_Score, A3_Score, A4_Score, A5_Score, A6_Score, A7_Score, A9_Score, A10_Score, ethnicity, contry_of_res}	100	10
14	{A1_Score, A2_Score, A3_Score, A4_Score, A5_Score, A6_Score, A7_Score, A8_Score, A9_Score, ethnicity}	100	10
15	{A1_Score, A2_Score, A3_Score, A4_Score, A6_Score, A7_Score, A9_Score, A10_Score, age, ethnicity}	100	10
16	{A1_Score, A2_Score, A3_Score, A4_Score, A5_Score, A6_Score, A9_Score, age, ethnicity, contry_of_res}	100	10
17	{A2_Score, A4_Score, A5_Score, A6_Score, A7_Score, A9_Score, A10_Score, gender, ethnicity, contry_of_res}	100	10
18	{A1_Score, A2_Score, A3_Score, A4_Score, A5_Score, A6_Score, A7_Score, A8_Score, A9_Score, contry_of_res}	100	10
19	{A2_Score, A3_Score, A4_Score, A5_Score, A6_Score, A7_Score, A8_Score, A9_Score, gender, contry_of_res}	100	10
20	{A1_Score, A6_Score, A7_Score, A8_Score, A9_Score, A10_Score, age, gender, ethnicity, contry_of_res}	100	10
21	{A1_Score, A4_Score, A6_Score, A8_Score, A10_Score, age, gender, ethnicity, Jaundice, contry_of_res}	100	10
22	{A1_Score, A2_Score, A4_Score, A6_Score, A7_Score, A10_Score, age, gender, ethnicity, Jaundice, contry_of_res}	100	11

Generation of reducts: In rough set theory, attribute reduction is achieved through generation of reducts, which is an essential feature in the rough set application to data mining. A reduct of an attribute set describes a minimum set of attributes representing the various classes of the data set. The objective of rough set theory is to derive minimal decision rule using reducts. The reducts are then used on the autism data set to derive the decision rules. These rules can then be used to categorize new instances of data and predict the unknown class label. The Rosetta tool has various algorithms based on heuristics for the formation of reducts. In this study, attribute reduction and classification are carried out using Johnson's algorithm and Genetic algorithm. The reduct generated by the algorithms for seed value of 4 are shown in Figures 3 and 4, respectively.

Generation of Decision Rules: Rough set methods generate rules which are used for predicting the output. The reduct algorithms in the Rosetta tool also generate IF_THEN rules consisting of predictor variables and a decision descriptor variable, which can be used for classification. A sample of the rules generated by Johnson's algorithm is shown in Figure 5. To determine the strength of the classifiers, they are validated on the training set and the final model is tested on the test set. Both Johnson's and genetic algorithms are applied to the training sets resulting from different training and test split. Figure 6 shows the performance of these algorithms for seed values ranging from 0.1 to 0.9.

Figure 5: Rules Created By The Johnson's Algorithm (Sample)

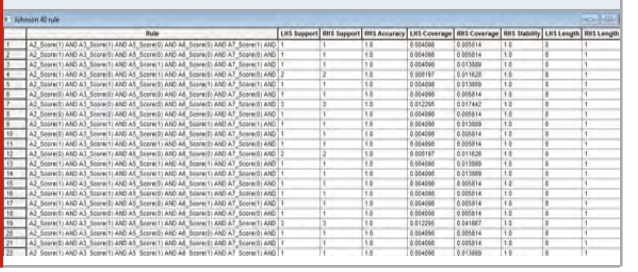
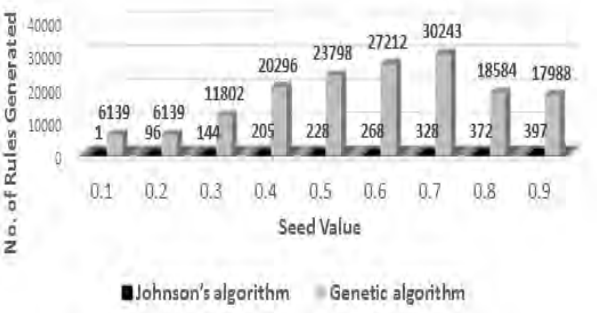


Figure 6: Performance of the Algorithms On The Autistic Data Set



The rules contain 15 predictor variables that predict the outcome variable 'ASD'. It can be seen from Figure 6, the total rules generated by Johnson's, and genetic algorithms were 144 and 11802 respectively. The total rules generated by the genetic algorithm is more than that of Johnson's algorithm for the various sample sizes. This is because GA generated a greater number of reducts.

Evaluation of the Proposed Model

Classification accuracy: The performance of the classifiers

is measured using accuracy, which is displayed using confusion matrix (Witten et al. 2011). The confusion matrix shows the number of correct and incorrect predictions given by the rules on the test data. In the binary classification, the objects in the test set will belong either to the positive decision class (Yes) or to the negative decision class (No). The confusion matrix generated by the Rosetta tool for genetic algorithm with seed value six is depicted in Figure 7.

Figure 7: Confusion Matrix For 40% Test Samples Using Genetic Algorithm

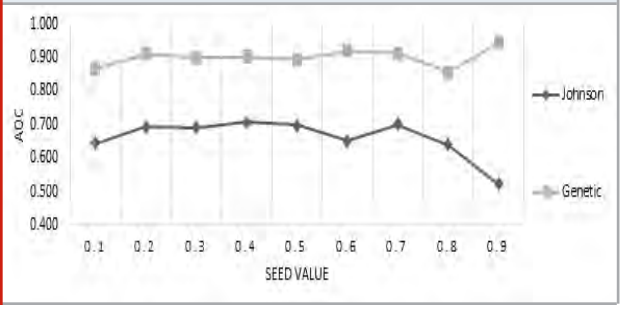
genetic 40

		Predicted		
		NO	YES	
Actual	NO	232	25	0.902724
	YES	10	98	0.907407
		0.958678	0.796748	0.90411
ROC	Class	YES		
	Area	0.920648		
	Std. error	0.01861		
	Thr. (0, 1)	0.5		
	Thr. acc.	0.432		

Table 1. Classification Accuracy % Of Rough Set Algorithms

Seed Value	Johnson's algorithm	Genetic algorithm
0.1	0.047	0.803
0.2	0.549	0.893
0.3	0.585	0.874
0.4	0.594	0.865
0.5	0.595	0.868
0.6	0.518	0.904
0.7	0.589	0.899
0.8	0.507	0.854
0.9	0.332	0.900

Figure 8: Auc Comparison Between The Classifiers



It can be seen that the accuracy for classifying 40% test samples using the genetic algorithm is 90.4%. Out of 365 samples in the test data set, 330 samples were classified correctly, and only a few samples were not classified correctly. The accuracy percentage of the two algorithms are given in Table 1. It is seen that the genetic algorithm has the highest accuracy among the two algorithms for all the seed values.

Receiver Operating Characteristic (ROC) Curves: Receiver Operating Characteristic (ROC) curves are helpful for evaluating the accuracy of predicted values (Fawcett 2004). It is a graphical representation for assessing the predicted and actual values in a classifier model. It is used to measure the performance of a test. ROC curves examine the relationship between sensitivity and specificity. The Area Under Curve (AUC) provides a measure of bias that how well a classifier can distinguish objects in class a label. AUC is normally recognized as the best measure for assessing the performance of classifiers (Øhrn 1999). The AUC values obtained for both the algorithms are shown in Figure 8. It is seen from the above Figure 8 that the genetic algorithm has greater AUC value is for all the seed values and it outperforms Johnson's algorithm. The dataset sizes that result in prediction models with the largest area under their ROC curves is selected as the better classification method for autism data.

CONCLUSION AND FUTURE WORK

A rough set model using ROSETTA tool kit has been applied on the UCI autism adult data set for predicting the presence or absence of ASD. Johnson's and genetic algorithms were used for classification task and the results reveal that the genetic algorithm has a superior performance over Johnson's algorithm with classification accuracy and ROC curve measures.

REFERENCES

Cao Y Liu S Zhang L Qin J Wang J and Tang K (2006) Prediction of Protein Structural Class with Rough Sets, BMC Bioinformatics, pp.7-20.

Daubie M Levecq P and Meskens N (2002) A Comparison of Rough Sets and Recursive Partitioning Induction Approaches: An Application to Commercial Loans, International Transactions in Operational Research, pp.681-694.

Dua D and Graff C (2019) UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Dubois D and Prade H (1991) Foreword In: Rough Sets - Theoretical Aspects of Reasoning about Data, by Z. Pawlak. Kluwer, Dordrecht, Netherlands.

Fawcett T (2004) ROC Graphs: Notes and Practical Considerations for Researchers, Kluwer Academic Publishers.

Han J and Kamber M (2001) Data Mining Concepts and Techniques, Morgan Kaufmann Publishers, ISBN 1-55860-489-8

Hassanien AE and Ali JMH (2004) Rough Set Approach for Generation of Classification Rules of Breast Cancer Data, INFORMATICA, Vol.15, No.1, pp.23-38.

Jalaja Jayalakshmi V Geetha V and Vivek R (2019) Classification of Autism Spectrum Disorder Data using Machine Learning Techniques, International Journal of

Engineering and Advanced Technology (IJEAT), ISSN: 2249 – 8958, Volume-8 Issue-6S, pp. 565-569.

Mak B and Munakata T (2002) Rule Extraction from Expert Heuristics: A Comparative Study of Rough Sets with Neural Networks and ID3, European Journal of Operational Research, Vol.136, pp.212–229.

Mitra S Mitra M. and Chaudhuri BB (2006) An Approach to Rough Set Based Disease Inference Engine for ECG Classification", Rough Sets and Current Trends in Computing, Proceedings of 5th International Conference, RSCRC, pp.400-406

Nandakumar GS Geetha V Surendiran B and Thangasamy S (2014) A Rough Set Based Classification Model for Grading in Adaptive E-Assessment, International Reviews on Computers & Software, Vol.9, No.7, pp.1169 -1177.

Øhrn A (1999) Discernibility and Rough Sets in Medicine: Tools and Applications, PhD Thesis, Norwegian University of Science and Technology, Trondheim, Norway.

Pawlak Z (1998) Reasoning about Data – A Rough Set Perspective, RSCTC'98, Lecture Notes in Artificial Intelligence, Vol.1424, pp.25-34.

Perzyk M. and Soroczynski A (2010) Comparative Study of Decision Trees and Rough Sets Theory as Knowledge Extraction Tools for Design and Control of Industrial Processes, World Academy of Science, Engineering and Technology, pp.84- 90.

Reidmiller M (1994) Advanced Supervised Learning in Multi-layer Perceptions from Back Propagation to Adaptive Learning Algorithms, Computer Standards Interfaces, pp.265-278.

Revett K (2005) A Rough Set Based Classifier for Primary Biliary Cirrhosis, International Conference on Computers as a Tool, pp.1128-1131.

Witten H Frank E and Hall MA (2011) Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann Publishers.

Biosc.Biotech.Res.Comm. Special Issue Vol 13 No 11 (2020) Pp-99-104

BBRC
Bioscience Biotechnology Research Communications

A Study on the Effectiveness of Machine Learning Algorithms in Early Prediction of Diabetics among Patients

R.K. Kavitha[1]* and W. Jai Singh[2]
[1,2]*Department of Computer Applications, Kumaraguru College of Technology, Coimbatore, India*

ABSTRACT

Nowadays healthcare industry seems to generate enormous data which when analysed using appropriate machine learning algorithms and tools shall provide greater insights of it. Such analysis helps to discover unusual and difficult to diagnose diseases at an early stage thus promising an increased success rate of curing such diseases and reduced medical expenses. This investigation aims to construct a model which will be able to foresee the chances of occurrence of diabetes among patients with greatest precision. Among the various machine learning algorithms which can be used for classifying data, three were chosen for this study and are listed as follows: Decision Tree, Naïve Bayes and Multilayer Perceptron. All the three techniques were applied on the diabetic data set and their performances were analysed using various metrics. Also, the results were compared by varying the k-fold values. Early stage diabetes risk prediction dataset obtained from UCI machine learning repository was utilized in this research work. This study has positively displayed the ability of the predicting patients with early diabetic risks in a large dataset. Among the three classification techniques, multilayer perceptron seems to classify the patient as diabetic or not with a higher degree of accuracy and with a chosen k-fold value of five and eight.

KEY WORDS: CLASSIFICATION, EARLY DIABETICS, MACHINE LEARNING, MEDICAL DIAGNOSIS, PREDICTION.

INTRODUCTION

Throughout the world, many people are being affected by Diabetes at an early age (Mohemaiti et al., 2017). Diabetes is considered as a chronic disease that affects the insulin production in humans thus affecting the entire body metabolism (Choubey et al., 2017). A spike in the blood glucose levels is observed because of this disease (Georga et al., 2013). Diabetes can damage vital parts of the human body. Increase in thirst and hunger, repeated urination are a few symptoms caused due to high blood glucose levels which may lead to complications if not treated at an early stage. Factors such as weight, height, insulin levels and genetic factors may cause diabetes in humans. Apart from the listed factors, the glucose concentration in the blood is considered as a major reason for this disease. The quick detection of the symptoms remains the only solution to reduce the risk of stroke, blood pressure, kidney diseases, foot, eye, and skin complications. With proper treatment and suggested lifestyle changes, several people with diabetes will be able to avoid or postpone the onset of complications.

Machine learning in the field of healthcare is gaining popularity nowadays and used for clinical decision making (Pradhan et al., 2012). Several machine learning algorithms can be employed to examine medical images to discover abnormalities, classify data into various classes based on several features and predict diseases at a very early stage. Classification is a task that uses machine learning algorithms to designate a class label to the data set. Classification is one of the most essential

ARTICLE INFORMATION

**Corresponding Author: kavitha.rk.mca@kct.ac.in*
Received 12th Oct 2020 Accepted after revision 9th Dec 2020
Print ISSN: 0974-6455 Online ISSN: 2321-4007 CODEN: BBRCBA

Thomson Reuters ISI Web of Science Clarivate Analytics USA and Crossref Indexed Journal

NAAS Journal Score 2020 (4.31) SJIF: 2020 (7.728)

Online Contents Available at: http//www.bbrc.in/
Doi: http://dx.doi.org/10.21786/bbrc/13.11/22

aspects of supervised learning. Several research studies have been conducted in various parts of the world to predict diseases using classification algorithms like SVM, Naive Bayes, J48, Decision Tree and so on (Kavakiotis et al., 2017; Vijayan 2015). This research focuses on predicting diabetics at an early stage by classifying the patients as diabetic or not based on the presence of several symptoms. Decision Tree, Naïve Bayes and Multilayer Perceptron algorithms were used in this study for early diabetic prediction. The performance of these algorithms was studied, and comparisons were drawn on various measures.

MATERIAL AND METHODS

Machine learning is a technique which when applied on medical datasets helps to detect and diagnose diseases in an effective manner (Dhomse Kanchan 2016). Diabetes, a universal disease which has created a serious impact on the society thus making machine learning as a priority in the medical research field. Several researchers have done a significant work in this field. Factors such as blood glucose levels, age, blood pressure, insulin, body mass index and skin thickness were used in a study to predict diabetes disease (Tarik et al., 2016). The contribution of adaboost and Bagging ensemble machine learning methods along with the support of decision tree technique to predict a patient as diabetic or non-diabetic was examined in a study (Perveen et al., 2016). The study has confirmed that adaboost technique performed better than the others. A ten-fold cross validation was performed on the data set, techniques like Logistic Regression, Naïve Bayes, and SVM was used in a study.

Figure 1: Workflow of Proposed model

Table 1. Description of the Dataset

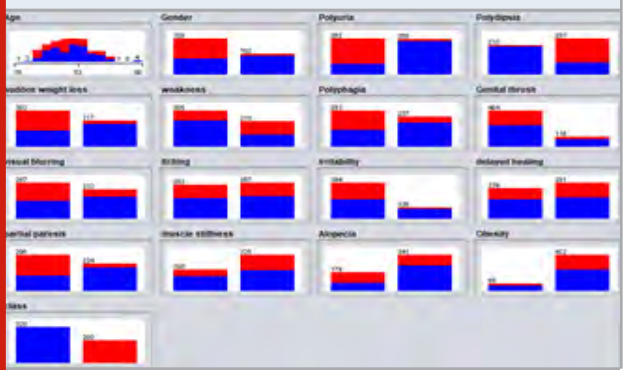
Database Name	Total Attributes	Total Instances
Early stage diabetes risk prediction dataset	17	520

It was concluded that SVM gave considerable accuracy with performance when compared to other methods (Kavakiotis 2017). A model that used machine learning decision tree concept was devised for diabetes prediction with the aim of predicting diabetes among age group of people (Orabi et al., 2016). Acquired results were found to be satisfactory in forecasting diabetes at a certain age, with better accuracy applying Decision tree (Rokach 2008). Another researcher studied the performance of KNN and DISKR in reducing the storage space and concluded that outlier removal has increased both performance and accuracy of model (Song et al., 2017). Yet another study claimed that Naive Bayes provided better accuracy rate of 77.01% in predicting diabetes (Rani 2016). Various popular machine learning classification methods such as Logistic Regression, Decision Tree, Naive Bayes as well as Artificial Neural Networks were used in classifying the diabetic risk among patients. To improve the strength of the proposed study, Bagging and Boosting techniques were employed. It was concluded that Random Forest algorithm displayed ideal results among the various other algorithms (Nongyao 2014).

Table 2. Attribute Description

Attribute Description	Possible Values/range
Age	20-65
Sex	Male, Female
Polyuria	Yes, No
Polydipsia	Yes, No
Sudden weight loss	Yes, No
weakness	Yes, No
Polyphagia	Yes, No
Genital thrush	Yes, No
Visual blurring	Yes, No
Itching	Yes, No
Irritability	Yes, No
Delayed healing	Yes, No
Partial paresis	Yes, No
Muscle stiffness	Yes, No
Alopecia	Yes, No
Obesity	Yes, No
Class	Positive, Negative

Figure 2: Overview of the dataset



Proposed Work: The suggested research workflow is shown in figure 1. The purpose of this study is to propose a classification model which can predict early stage diabetics among patients. The dataset was preprocessed to enhance the quality and to transform raw data into an understandable and readable format. Noise removal, handling of missing data was carried out. K-fold cross-validation technique helped to split data as training and test sets. Data set was divided as k successive folds and every fold will be used once for validation whereas the left behind k - 1 folds is used as training set. K-fold values of 5,6,7 and 8 were used in the study.

Dataset: The suggested methodology is assessed on Early stage diabetes risk prediction dataset obtained from UCI machine learning repository. Dataset comprised of health details of 520 samples, including the sign and symptoms particulars of freshly diabetic or would be diabetic persons. The dataset contained 17 attributes whose description and possible values are shown in table 1 and 2. The number of patients classified as diabetic positive and diabetic negative for each attribute is shown in figure 2.

Classification Techniques:

Decision Tree: Decision Tree is considered as a supervised machine learning technique employed to classify data. The major intention of utilizing the technique here is to predict target class with the help of decision rules brought from earlier data (Agarwal et al., 2016). Internodes and nodes are used to predict and classify data. The root nodes are used to classify the various features found in the data. There may be branches appearing from the root node and the leaf nodes appearing from it classifies data. Information gain will be calculated for the attributes at each stage and the highest value of it will be used by the decision tree to choose each node (Rokach 2008).

Table 3. Accuracy measures

Performance Metric used	Meaning	Calculation
Accuracy	Lends overall classifier effectiveness, reveals how many times the classifier provided correct result in the given sample.	A=(TP+TN) / (Total no of samples)
Precision	Ratio of true positives to the total number of prediction as positives.	P = TP / (TP+ FP)
Recall	Ratio of true positives to the toral number of actual positives.	R =TP / (TP+FN)
F-Measure	Harmonic mean of precision and recall, gives equal importance to precision and recall.	F=2/(1/recall + 1/precision)
ROC	Graphic plot comparing classifier's True Positive rates, False Positive rates.	Not Applicable

Table 4. Classifier: J48 – Decision Tree

S. No	No. of K-fold	True Positive Rate	False Positive Rate	Precision value	Recall value	ROC Area
1	5	0.944	0.055	0.965	0.944	0.944
2	6	0.952	0.043	0.953	0.952	0.962
3	7	0.954	0.048	0.954	0.954	0.967
4	8	0.948	0.053	0.948	0.948	0.962

Table 5. Classifier: Naïve Bayes

S. No	No. of K-fold	True Positive Rate	False Positive Rate	Precision value	Recall value	ROC Area
1	5	0.865	0.124	0.873	0.865	0.944
2	6	0.875	0.116	0.881	0.875	0.947
3	7	0.869	0.119	0.877	0.869	0.946
4	8	0.877	0.114	0.883	0.877	0.947

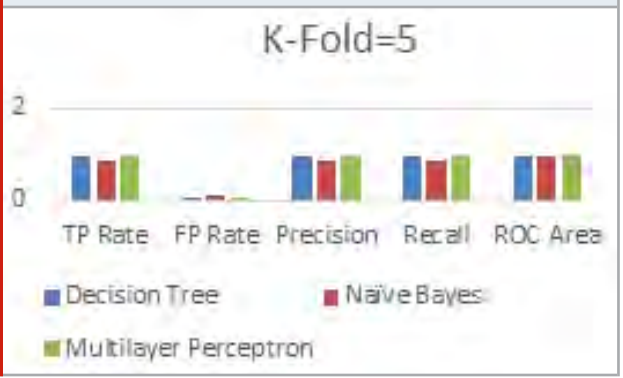
Table 6. Classifier: Multilayer Perceptron

S. No	No. of K-fold	True Positive Rate	False Positive Rate	Precision value	ROC Area
1	5	0.960	0.036	0.961	0.960
2	6	0.950	0.048	0.951	0.950
3	7	0.962	0.039	0.962	0.962
4	8	0.962	0.037	0.962	0.962

Naïve Bayes: It can be understood that the entire features in Naive Bayes classification technique are not dependent and at the same time not related to each other. Here, the importance of a feature observed in a class will not affect other features. Being built on conditional probability, this technique is believed to be powerful in classifying data. It operates well on imbalanced dataset and data with missing values. The popular Bayes Theorem was employed by this classification technique (Mani et al., 2012).

Multilayer Perceptron: A multilayer perceptron (MLP) normally combines along with further perceptrons, piled in numerous layers with the idea of solving complicated problems. In this technique, more than one linear layer can be present (Rana et al., 2018). In a three-layer network, initial layer named as input layer, final called as output layer and the middle one as hidden layer. Normally, data is supplied into the input layer and obtained as output from the last layer. The hidden layer count can be increased as required in order to make the model more complex according the mission. The multilayer perceptron thus offers a nonlinear mapping amongst the input vector and output vector. The steps to be followed by perceptron are (1) Take the inputs, multiply by their weights and computes their sum (2) Adds a bias factor, the number one multiplied by a weight and (3) Feeds the sum all the way through activation function.

Figure 3: Classifier performance for K-Fold 5



Performance Metrics: To determine the implementation of classification technique, several metrics were used in the study. Accuracy, F-Measure, Recall, Precision and Receiver Operating Curve (ROC) measures were utilized for data classification. The details of the accuracy measures are exhibited in table 3.

Figure 4: Classifier performance for K-Fold 6

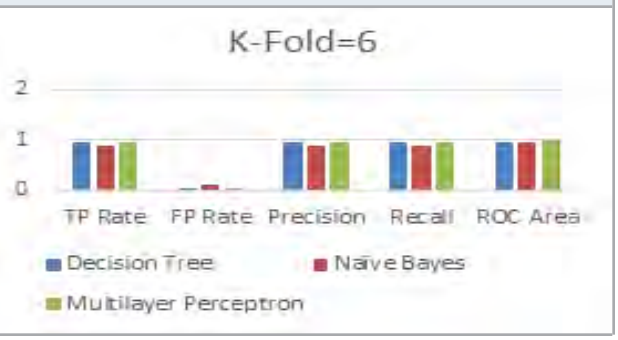


Figure 5: Classifier performance for K-Fold 7

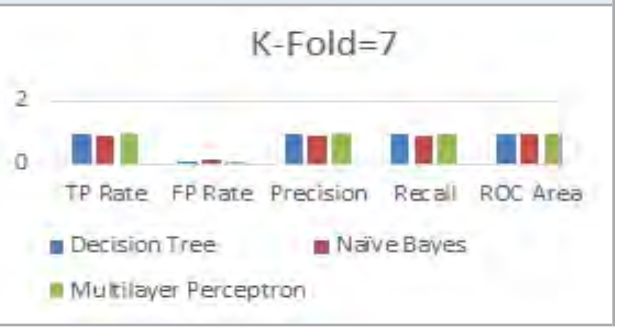
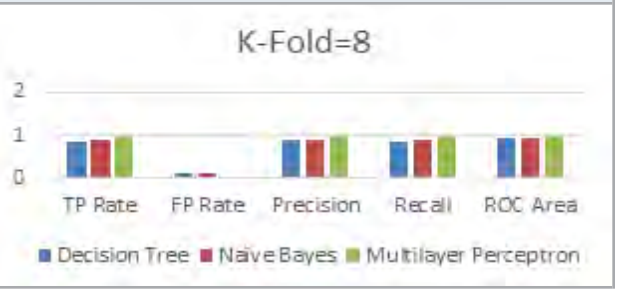


Figure 6: Classifier performance for K-Fold 8



RESULTS AND DISCUSSION

Tables 4, 5 and 6 represents the performance of three classification algorithms used in the study namely Decision Tree, Naïve Bayes, and Multilayer Perceptron. Values of different metrics for the K-fold values of five,

six, seven and eight are presented in the table. It can be interpreted from the results that the Multilayer Perceptron algorithm can predict the early diabetes with maximum precision, TP rate and ROC area when compared to other algorithms. Further, the performances of the three classifiers based on k-fold values are plotted as a graph in figures 3,4,5 and 6.

Table 7. Comparison of classification techniques with error rates

S. No	Model	No. of K-fold	In-Correctly Classified	Instances Correctly Classified	Kappa Statistic	Mean Absolute error value	Root mean squared error	Relative absolute error (%)
1	J48 – Decision Tree	5	491 (94.4%)	29 (5.5%)	0.883	0.077	0.232	16.28
2	J48 – Decision Tree	6	495 (95.1%)	25 (4.8%)	0.899	0.063	0.214	13.4
3	J48 – Decision Tree	7	496 (95.38%)	24 (4.6%)	0.902	0.059	0.208	12.65
4	J48 – Decision Tree	8	493 (94.8%)	27(5.1%)	0.890	0.067	0.223	14.31
5	Naïve Bayes	5	450 (86.5%)	70 (13.4%)	0.722	0.150	0.320	31.8
6	Naïve Bayes	6	455 87.5%	65 12.5%	0.742	0.147	0.316	31.0
7	Naïve Bayes	7	452 87%	68 13%	0.730	0.149	0.319	31.6
8	Naïve Bayes	8	456 87.6%	64 12.3%	0.745	0.147	0.317	31.0
9	Multilayer Perceptron	5	499 96%	21 4%	0.915	0.044	0.173	9.4
10	Multilayer Perceptron	6	494 95%	26 5%	0.895	0.053	0.193	11.21
11	Multilayer Perceptron	7	500 96%	20 4%	0.9191	0.045	0.179	9.65
12	Multilayer Perceptron	8	500 96%	20 4%	0.919	0.043	0.180	9.18

A comparison of the classifiers with error rates are presented in the table 7. It may be inferred that Multilayer Perceptron algorithm performs well in correctly classifying the diabetic patients when compared to Decision Tree and naïve Bayes algorithms. The mean absolute error, root mean squared and relative absolute error rates for multilayer perceptron is comparatively low than the other algorithms used in the study. Also, it can be understood that the multilayer perceptron algorithm works better for chosen k-fold values of 5 and 8.

CONCLUSION

One of the crucial needs nowadays is the detection of diabetes at its initial phase. The proposed model helps to detect early diabetes among patients with high accuracy. The prediction outcomes of three machine learning classification algorithms were analysed considering various metrics and measures. Experiments were performed using early stage diabetes risk prediction dataset taken from UCI machine learning repository. The study findings suggest that Multilayer Perceptron algorithm outperforms the other two algorithms in classifying the data set as diabetic with more precision and less error rates. Also, this algorithm can be used to predict various other diseases in coming days.

Conflict of Interest: The authors declare no conflicts of interest.

REFERENCES

Agarwal V, Podchiyska T, Banda J.M et al (2016) Learning statistical models of phenotypes using noisy labeled training data Journal of the American Medical Informatics Association Vol 23 No 6 Pages 1166–1173

Choubey D.K, Paul S, Kumar S (2017) Classification of Pima Indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection Communication and Computing Systems Proceedings of the International Conference on Communication and Computing System (ICCCS 2016) Pages 451–455

Dhomse Kanchan B (2016) Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication, IEEE Pages 5–10

Georga E.I, Protopappas V.C, Ardig`o D et al (2013) Multivariate Prediction of Subcutaneous Glucose Concentration in Type 1 Diabetes Patients Based on Support Vector Regression IEEE Journal of Biomedical Health Informatics Vol. 17 No 1 Pages 71–81

Kavakiotis I, Tsave O, Salifoglou A et al (2017) Machine Learning and Data Mining Methods in Diabetes Research Computational and Structural Biotechnology Journal Vol 15, Pages 104–116

Mani S, Chen Y, Elasy T (2012) Type 2 Diabetes Risk Forecasting from EMR Data using Machine Learning AMIA Annual Symposium Proceedings Pages 606–615

Mohemaiti P, Keyoumu Y et al (2017) Current situation and related risk factors of elderly type 2 diabetes mellitus with coronary heart disease in Hangzhou road community of the New Urban Area of Urumqi Chinese Journal of Gerontology Vol 37 No 21 Pages

5422–5424
Nongyao N, Punnee S (2014) Ensemble Learning Model for Diabetes Classification Advanced Materials Research Vol 931-932 Pages 1427-1431
Orabi K.M, Kamal, Y.M, Rabah, T.M (2016) Early Predictive System for Diabetes Mellitus Disease Industrial Conference on Data Mining Springer Pages 420-427
Perveen S, Shahbaz M, Guergachi A et al (2016) Performance Analysis of Data Mining Classification Techniques to Predict Diabetes Procedia Computer Science Vol 82 Pages 115–12
Pradhan,P.M.A,Bamnote G.R.,Tribhuvan V et al (2012) A Genetic Programming Approach for Detection of Diabetes. International Journal of Computational Engineering Research Pages 91–94
Rana A, Singh Rawat A, Bijalwan et al (2018), "Application of Multi-Layer Perceptron Artificial Neural Network in the Diagnosis System: A Systematic Review International Conference on Research in Intelligent and Computing in Engineering (RICE) San Salvador Pages 1-6
Rani A, Swarupa, Jyothi S (2016) Performance analysis of classification algorithms under different datasets In Computing for Sustainable Global Development (INDIACom) Third International IEEE Conference Pages 1584-1589
Rokach L, Maimon Z.O (2008) Data mining with decision trees: theory and applications Vol 13 World Scientific Publishing Co Toh Tuck Singapore
Song, Yunsheng, Jiye Liang et al (2017) An efficient instance selection algorithm for k nearest neighbour regression Neurocomputing Pages 26-34
Tarik A, Rashid,S.M.A, Abdullah,R.M. (2016) An Intelligent Approach for Diabetes Classification Prediction and Description Advances in Intelligent Systems and Computing Pages 323–335
Vijayan,V.V, Anjali,C (2015) Prediction and diagnosis of diabetes mellitus A machine learning approach IEEE Recent Advances in Intelligent Computational Systems (RAICS) Pages 122–127

Biosc.Biotech.Res.Comm. Special Issue Vol 13 No 11 (2020) Pp-105-108

BBRC
Bioscience Biotechnology Research Communications

Analysis of Microarray Gene Expression Data Using Various Feature Selection and Classification Techniques

W. Jai Singh[1]* and R. K. Kavitha[2]
[1&2]*Assistant Professor (SRG), Department of Computer Applications*
Kumaraguru College of Technology, Coimbatore, India

ABSTRACT

The prediction and diagnosis of the cancer disease can be augmented by applying classification techniques on the microarray-based gene profiling data. It is observed that a massive data will be generated due to the regular changes that happen in the cancer producing genes among humans. Along with the vast expression of genes, only a minor fraction of it are substantially articulated. By performing analysis of microarray data, the cancer-causing genes can be identified with higher accuracy. Normally, feature selection will be performed on the data followed by the classification process. The objective of this manuscript is to select the meaningful gene features of DNA microarray data with the help of ANOVA technique thus displaying an enhanced performance of algorithms like Ada Boost, Neural Networks and Random Forest. To evaluate the suggested method, it is planned to lessen the gene count from breast and leukemia obtained from DNA microarray dataset. Investigational results on the above-mentioned datasets reveals that the gene picked by the suggested methodology displays an improved classification accuracy of Neural Networks. This work tries to draw a comparison between Neural Networks and ANOVA techniques along with classical feature selection method and more classifiers. It can be concluded that ANOVA feature selection method along with neural networks offers improved accuracy of classification. Also, the ROC reveals the excellent subclass of genes with a better classification accuracy.

KEY WORDS: ANOVA, FAST CORRELATION BASED FILTER, GENE EXPRESSION, INFORMATION GAIN, MICROARRAY DATA, NEURAL NETWORKS.

INTRODUCTION

Currently, in the field of biomedical and clinical research, the analysis of gene expression with the help of microarrays is gaining popularity. Latest developments in the field of DNA microarray provided a way to examine and assess the communication levels of several microarray data genes at the same time thus allowing generation of huge microarray data (Fan et. al. 2009). Microarrays have been extensively used in several aspects of biomedical investigations since this technique enabled researchers to perform enormous testing on gene patterns (Vilda et. al 2006). Availability of more dimensions in the microarray gene data makes it possible to assess every single gene in a particular environment on several categories of tissues (Zheng et.al. 2006). The property of extracting valuable data from microarray data analysis has drawn the attention of researchers from various field.

Also, this method poses the challenge of translating such data on the way to obtain a clear-cut understanding of the biological procedures and the diseases occurring in humans (Peng et.al. 2006). The highlights of this work are listed as follows: (i) Several feature selection approaches centered on statistical tests that helps to handle data of high dimensions are suggested. Information Gain, ANOVA and Correlation filter tests were used to identify suitable

ARTICLE INFORMATION

*Corresponding Author: jaisingh.w.mca@kct.ac.in
Received 15th Oct 2020 Accepted after revision 7th Dec 2020
Print ISSN: 0974-6455 Online ISSN: 2321-4007 CODEN: BBRCBA

Thomson Reuters ISI Web of Science Clarivate Analytics USA and Crossref Indexed Journal

NAAS Journal Score 2020 (4.31) SJIF: 2020 (7.728)

*Online Contents Available at: http//www.bbrc.in/
Doi: http://dx.doi.org/10.21786/bbrc/13.11/23*

features. (ii) Once the features were chosen, several methods namely Ada Boost, Random forest and Neural Networks were used for dataset classification.

MATERIAL AND METHODS

Related Work: A novel way to diagnose cancer disease is made possible using the enormous microarray gene expression technique. With applying classification algorithms on the gene data, it is possible to identify the most useful and important genes which causes cancer disease. Treatment can be started at the appropriate time to the patients with the help of gene identification. One of the significant steps for classification is to decrease the gene data dimension. This can be achieved with the help of feature selection method. A feature selection framework that blends gene analysis with several other methods was discussed by Tan et al (Tan et.al 2008). It was decided that by blending various techniques, effective results can be obtained rather than using a single component algorithm. A MRMR minimum redundancy maximum relevance feature selection framework was developed by Peng (Ding et. al. 2005) which removed the data redundancies present in the gene data. Huang and Chow (Huang et.al. 2005) presented a meaningful gene attributes by using the information gain on a predefined class labelled data.

Mutual information was employed by Zhang et al (Zhang et. al. 2009) in the classification problems with multi labels and it was demonstrated that MIM displayed an improved classification accuracy. François et al (2007) by using the feature selection and feature transformation in his research work, showed an improvement in the strength of forward feature selection method. A greedy feature selection process employing mutual information theory was recommended by Hoque et al (2014). The approach fuses feature–feature mutual information and feature–class mutual information towards achieving an ideal sub class of features thus minimizing the redundancy and maximising the feature relevance. Sha-Sha et al (2014) developed unified MIM into a system which used cloud computing, to classify gene expression data. This way of data classification seemed to improve efficiency of the program with identical classification accuracy.

Proposed work: The existence of many trivial and inappropriate features in the data set lowers the analysis quality of cancer like disease. To overcome this problem, this work proposes to explore the dataset in a right perspective. A new way of microarray data classification is presented in this section. The approach consisted of a two-phase execution. Initially, the data input was pre-processed with the help of techniques like identifying missing data, normalising data, and performing feature selection. Subsequently, classification techniques were utilised for microarray data set classification into cancer or non-cancer samples. The steps followed is as given below:

(a) Collection of data: The study dataset was extracted from National Center of Biotechnology Information (NCBI GEO, http://www.ncbi. nlm.nih.gov/gds/). This data was provided as input to the proposed model for classification.

(b) Handling Missing data, data set normalization: Computation of mean value was used to fill the missing data in the dataset. Min–Max normalization was used to standardise the input feature values within the range value [0:1] (Kumar et.al. 2014).

(c) Dataset division: The dataset is split up into two types namely a training set and testing set.

(d) Feature selection: Tests namely Information Gain, ANOVA and Correlation filter were used to pick features with high-ranking significant values thereby reducing the data dimensionality.

(e) Classifier design: Ada Boost, Neural Networks and Random forest techniques were used to categorize microarray dataset.

(f) Model testing: For a chosen K fold value, the classifier model was tested with the help of test dataset and subsequently, classifier performance was assessed by referring to achieved accuracy.

Selection of features using ANOVA: Analysis of Variance (ANOVA) is a parametric statistical hypothesis test used for verifying whether the mean value of two or more data samples arrive from a similar distribution or not. F-test which belongs to a set of statistical tests is used to calculate the ratio among variance values obtained from two samples. The ANOVA technique is a kind of F-statistic hence called as ANOVA f-test. Analysis of Variance test is used when one of the variables under consideration is of numeric type and another one is categorical in nature. Results of the above test can be utilized for selecting features. Also, this helps in removing the target variable independent features.

Feature Selection Method with IG: Information Gain (IG) is an evaluation technique based on entropy which is commonly used in supervised learning research. IG can be used for categorical class label by using the various attribute values. IG is considered to be an essential measure for feature ranking. Provided the entropy as a criteria for impurity in training dataset S, a measure indicating further details about Y supplied by X can be defined which is known as IG.

Feature Selection Method employing Fast Correlation Based Filter (FCBF) Two phases of FCBF is Significance analysis and Redundancy analysis.
In Significance Analysis, correlation is generally applied to examine the relevance. The relevance r can be calculated using the following equation:

$$r = \frac{\sum_i (x_i - \overline{x_i})(y_i - \overline{y_i})}{\sqrt{\sum_i (x_i - \overline{x_i})^2}\sqrt{\sum_i (y_i - \overline{y_i})^2}}$$

Nevertheless, most systems in real world applications are observed to be non-linear. Correlation in a non-linear system can be determined with the help of Symmetrical Uncertainty (SU).

$$SU = 2\left[\frac{IG(X|Y)}{H(X)H(Y)}\right]$$

$$IG(X,Y) = H(X) - H(X|Y)$$

$$H(X) = -\sum_{i} P(x_i) log_2 P(x_i)$$

Where IG(X,Y) represent the Information Gain. Following the positioning of relevant features, FCBF eliminates redundant features from selected features cantered on SU amongst feature and class, and among feature and feature. This is known as redundancy analysis.

Artificial Neural Networks (ANN) for Classification: The way in which human brain understands and manages information processing paved the way for data classification using ANN. It has the capability to cope with complicated features in the data. This characteristic helps in generalising and predicting the future cases with higher accuracy. In ANN, neurons are arranged in numerous layers. The concealed layer count and the presence of neurons in every layer is determined by the problem complexity. In order to collect predictor variables as input in the form of a node, the input layer communicates with outside environment. This information will be transferred to first hidden layer which is further multiplied by associated weights. Finally, these multiplied values are summed up and supplied through a nonlinear transfer function. The values are scaled further, and an output is produced which resembles to the axon of the neuron. Cancer classifiers based on Neural network have been employed together with binary-class and multi-class problems. This helps to identify cancer or non-cancer samples, type of cancer, or the risk of survival.

Table 1. Classification Results using Breast and Leukemia dataset with all features

Classifier	AUC		Classification Accuracy	
	Breast Dataset	Leukemia Dataset	Breast Dataset	Leukemia Dataset
Adaboost	0.477	0.818	0.485	0.594
Neural Network	0.710	0.874	0.670	0.625
Random Forest	0.666	0.769	0.598	0.656

Experimental Results and Discussions: Breast and leukemia microarray gene data are used to evaluate the various classification techniques. The Breast dataset contains a total of 97 samples with 24481 number of genes. The leukemia dataset contains 72 samples with 7219 genes. The combination of ANOVA with Neural Networks are proposed. The performance of the classifier is evaluated using classification accuracy and AUC - Area Under the Curve. Table 1 to 4 shows the comparison results.

Table 2. Classification Results using Breast and Leukemia dataset with ANOVA

Classifier	AUC		Classification Accuracy	
	Breast Dataset	Leukemia Dataset	Breast Dataset	Leukemia Dataset
Adaboost	0.619	0.838	0.629	0.750
Neural Network	0.881	0.973	0.804	0.844
Random Forest	0.822	0.943	0.763	0.734

Table 3. Classification Results using Breast and Leukemia dataset with fast correlation filter

Classifier	AUC		Classification Accuracy	
	Breast Dataset	Leukemia Dataset	Breast Dataset	Leukemia Dataset
Adaboost	0.565	0.704	0.567	0.562
Neural Network	0.798	0.849	0.711	0.703
Random Forest	0.866	0.825	0.773	0.609

Table 4. Classification Results using Breast and Leukemia dataset with Information Gain

Classifier	AUC		Classification Accuracy	
	Breast Dataset	Leukemia Dataset	Breast Dataset	Leukemia Dataset
Adaboost	0.636	0.804	0.639	0.703
Neural Network	0.817	0.914	0.763	0.781
Random Forest	0.853	0.848	0.753	0.703

In Table 1, the Breast and Leukemia microarray gene expression data are categorised by Adaboost, Neural Networks and Random Forest with all gene attributes. In the Table 2, the attributes are selected by ANOVA with classification techniques is applied in the data set. In Table 3, fast correlation filter method is applied to select the features. In Table 4, Information Gain as a feature selection method. It can be seen from Tables 1 to 4, the ANOVA + Neural Networks perform better than other individual and combinational methods. Therefore, the proposed approach improves the classification performance of Breast and Leukemia microarray data. Figure 1 to 4 show the performance of the various methods and proposed method using ROC Curve. Figure 3 and 4 shows pictorial representation for the comparison

result of Breast and Leukemia cancer dataset. The peak of the graph shows the maximum classification accuracy of these dataset by using the ANOVA + Neural Networks.

CONCLUSION

In this paper, an innovative method of microarray data classification for cancer has been suggested. The proposed method includes dimensionality reduction of gene expression data with the help of ANOVA technique, Information Gain, Neural Network and Fast Correlation filter. Research results suggest that using ANOVA with Neural Networks proves to be useful and efficient for classification.

Conflicts of Interest: The authors declare no conflicts of interest.

REFERENCES

Ding C, H. Peng (2005), Minimum redundancy feature selection from microarray gene expression data, J. Bioinf. Comput. Biol. 3 (02), 185–205.

Fan L, K.-L. Poh, P. Zhou (2009), A sequential feature extraction approach for Naïve Bayes classification of microarray data. Expert Syst. Appl. 36, 9919–9923.

François D, F. Rossi , V. Wertz , M. Verleysen (2007), Resampling methods for parameter-free and robust feature selection with mutual information, Neurocomputing 70 (7), 1276–1288 .

Hoque N, D.K. Bhattacharyya , J.K. Kalita (2014) , MIFS-ND: a mutual information-based feature selection method, Exp. Syst. Appl. 41 (14), 6371–6385 .

Huang D, T.W. Chow (2005) , Effective feature selection scheme using mutual information, Neurocomputing 63, 325–343.

Kumar M, S. Kumar Rath (2014), Classification of microarray data using kernel fuzzy inference system, Int. Scholarly Res. Notices, http://dx.doi.org/10.1155/2014/769159.

Peng Y (2006), A novel ensemble machine learning for robust microarray data classification. Comput. Biol. Med. 36, 553–573.

Sha-Sha W, L. Hui-Juan , J. Wei , L. Chao (2014), A construction method of gene expression data based on information gain and extreme learning machine classifier on cloud platform, International Journal of Database Theory and Application 7 (2), 99–108 .

Tan F, X. Fu , Y. Zhang , A.G. Bourgeois (2008), A genetic algorithm-based method for feature subset selection, Soft Comput. 12 (2), 111–120.

Vilda P.G, F. Díaz, R. Martínez, R. Malutan, V. Rodellar, C.G. Puntonet (2006), Robust preprocessing of gene expression microarrays for independent component analysis. Independent Component Analysis and Blind Signal Separation, Springer, pp. 714–721.

Zhang M L, J.M. Peña (2009), V. Robles , Feature selection for multi-label naive Bayes classification, Inf. Sci. 179 (19), 3218–3229 .

Zheng C.H, D.-S. Huang, L. Shang (2006), Feature selection in independent component subspace for microarray data classification. Neurocomputing 69, 2407–2410.

Biosc.Biotech.Res.Comm. Special Issue Vol 13 No 11 (2020) Pp-109-115

BBRC
Bioscience Biotechnology Research Communications

Detection of Diseases in Sugarcane Using Image Processing Techniques

K.Thilagavathi[1*], K.Kavitha[2], R.Dhivya Praba[3], S.V.Arockia Joseph Arina[4] and R.C.Sahana[5]
[1]*Asst.Professor, Department of ECE, Kumaraguru College of Technology, Coimbatore, India*
[2]*Professor, Department of ECE, Kumaraguru College of Technology, Coimbatore, India*
[3]*Asst.Professor, Department of ECE, Kumaraguru College of Technology, Coimbatore, India*
[4,5]*UG Students, Department of ECE, Kumaraguru College of Technology, Coimbatore, India*

ABSTRACT

India is one of the largest producers of sugarcane and ranks second in the world. The cropping season and duration of sugarcane depends on the varieties and sowing time. The time taken for its maturity is between 12 and 18 months. With high sensitivity to the environment, it easily gets affected by numerous diseases and pests. If the affected plant is not identified and taken adequate measures at the right time, it will affect the whole yield. The present study focuses on detecting various diseases in sugarcane leaves using the image processing techniques and developing a web application for the farmers to detect the major diseases of sugarcane as well. The system collects the images of the leaves and processes by means of Adaptive Histogram Equalization (AHE) superseded by segmentation using k means clustering algorithm. Using Gray Level Co-occurrence Matrix (GLCM) and Principal Component Analysis (PCA), statistical characteristics such as variance, skewness, standard deviation, mean, and covariance are extracted. Finally, the detection and classification are implemented using Support Vector Machine (SVM) that results the average accuracy value of 95%. If any variety disease is identified, the required control measures are also suggested.

KEY WORDS: SUGARCANE, FEATURE EXTRACTION, DISEASE, CLASSIFICATION, SUPPORT VECTOR MACHINE.

INTRODUCTION

India stands first amidst the sugarcane growing countries worldwide in terms of area beforehand (3.93 m.ha) and second in production (167 m.t). Despite the increase in the substantial yield per hectare in our country, the productivity of sugarcane is still lower when compared to other countries. The long duration crop faces various abiotic problems like shortage of water, differences in temperature, floods, nutritional lag, and alkalinity. Fungal diseases such as brown spot, wilt, rust, red rot, and smut caused by photo plasma and viral diseases like sugarcane streak mosaic, sugarcane mosaic virus, yellow leaf syndrome affect the yield greatly. Pests such as sugarcane borer, white wooly aphid are also not the least responsibility for the reduced yield of sugar productivity. Being a prominent cash crop of India, it tops the list by serving feed for live stocks, as fuel and its stubble and roots as organic manure. These diseases limit the sugarcane production and bring forth heavy losses. It is reported to have more than fifty diseases in sugarcane leaves. The most destructive agents are fungi, bacteria, viruses, and nematodes [Viswanathan and Rao, 2011]. The causes, symptoms,transmission, and control measures of these severe diseases are explained below.

ARTICLE INFORMATION

*****Corresponding Author: thilagavathi.k.ece@kct.ac.in
Received 27th Oct 2020 Accepted after revision 14th Dec 2020
Print ISSN: 0974-6455 Online ISSN: 2321-4007 CODEN: BBRCBA

Thomson Reuters ISI Web of Science Clarivate Analytics USA and Crossref Indexed Journal

NAAS Journal Score 2020 (4.31) SJIF: 2020 (7.728)

Online Contents Available at: http//www.bbrc.in/
Doi: http://dx.doi.org/10.21786/bbrc/13.11/24

Eyespot: Helminthosporium sacchari causes the eyespot disease in sugarcane. It grows from 6 - 7 months and is more susceptible to the disease. There occurs lessions which resemble small and water soaked spots initially and protrudes deeper into the surrounding tissues. The use of Foliar's 0.3% Mancozeb or 0.2% copper oxychloride with 2 to 3 sprays over the duration of 10 to 15 days can minimise the disease. Spraying should be done based on the severity of the disease. [https://plantvillage.psu.edu/topics/sugarcane/infos]. Leaf Scald: Leaf Scald is another disease caused by Xanthomonas Albileneans. This disease may be controlled effectively by several methods. The disease resistant breeds can be approached from nurseries and are cultivated. These cuttings can be developed using sets which have undergone a cold soak/long hot-water treatment. After treatment with hot- water at 50°C for 3 hours, the cuttings are allowed to soak in cold flowing water for 48 hours. [https://plantvillage.psu.edu/topics/sugarcane/infos].

Yellow Leaf: The sugarcane yellow leaf is affected by the Sugarcane Yellow Leaf Virus (SCYLV).The yellowing extends from the leaf midrib across leaf blades during the monsoon season until a general yellowing of the leaves can be seen from a distance.[Devi Aruna,2016]. The most successful method for the control of sugarcane diseases is to boost varietal resistance breeds. The lack of knowledge in choosing the variety of genetic yellow leaf resistance breed makes the crop development difficult for this disease. [Viswanathan and Rao,2011].

Red Rot: A fungal disease caused by Colletotrichumfalcatum Wentis is called Red rot. The noticeable signs during severe crop infection are dull red to brown shades on the rind / nodal regions, pinkish porulation at the rind and leaf scars. The leaves can be treated for 15-20 minutes by dipping them in 0-1 percent carbendazim (Bavistin). This should be done before primary and general cane planting. Spraying should be performed immediately after the outbreak of the disease. [https://sugarcane.icar.gov.in/index.php/en/2014-04-28-11-31-50/major-diseases] Mosaic: The mosaic disease is caused by the Sugarcane Mosaic Virus (SCMV) / Sugarcane Streak Mosaic Virus (SCSMV). Roguing the infected plants to eliminate SCMV can been successful. Incorporating inoculum pressures, a control measure, are not expressive among the use of mosaic- free seed canes. Planting material thermo-therapy may result in SCMV-free among some plants. The creation and use of resistant clones dominated the mosaic in sugarcane for a long time. [Devi Aruna, 2016].

Ringspot: It is caused by the fungus Leptosphaeriasacchari. Tiny, elongated, oval-shaped spots turning from dark olive green to reddish-brown with thin yellow halos, are the initial signs of a ring-spot. Spots are often visible on the leaf sheaths and stems. Mostly the older leaf blades are become the most affected part of the plant. Soil amendments and calcium silicate slag produces a substantial positive effect on yield reducing the intensity of ring spots.

MATERIAL AND METHODS

The authors Al Hiary et al,2011, suggested an algorithm for the automatic detection and classification of leaf diseases in plants that are experimentally tested. The algorithm has four main phases. First, the algorithm identifies green colored pixels in the leaf. The identified green pixels have masked using the threshold computed by Otsu's method. Then the pixels of the infected clusters were identified and removed. Then k-means clustering, and Neural Networks (NNs) were implemented for clustering and classification of plant diseases. Dheeb et al.,2011, suggested a framework to identify the leaf/stem diseases. The algorithm was tested with the images of the plants distressed by six distinct forms of diseases. Various soft computing / image segmentation methods proposed by Vijai Singh and Misra AK (2017) for the classification of plant diseases are analysed in this paper. The authors also propose an algorithm that automatically classifies and detects plant diseases by means of image segmentation and classification approaches. Further, the performance characteristics of the suggested algorithm has been tested on the images of various infected plants. From the survey it has been understood that by combining different algorithms could increase the cognition rate.

The input leaf image is pre-processed primarily using AHE in the proposed work and pre-ceded by the segmentation using k-means clustering algorithm, which utilises rehearsal refinement to produce the final result. The statistical features are extracted using GLCM and PCA. Finally the detection and classification are implemented through SVM and it compares the extracted image to the reference image. The displayed results infer whether the leaf is affected. If any leaf is affected with any disease, the disease type is identified. The various steps involved in the proposed method are given in the flow diagram, in the figure 1. The images are processed beforehand the classification. All type of data given is conditioned using enhancing, displaying, and extracting information. The methodology for detection of diseases on the sugarcane leaf is done using MATLAB software [Sandesh 2017].

The input of the developed application is the image of the leaf and is processed using adaptive histogram equalization. Segmentation is carried out with k means clustering algo-rithm that produces tighter clusters, particularly when they are globular, than any other clus-ters. Then extraction is performed using Grey level co-occurrence matrix and Principal Component Analysis and finally the detection and classification is implemented using sup-port vector machine. The diseased and healthy leaves of sugarcane are placed in the database for android applications. So, the extracted image is compared with that reference image.

The results show whether the leaf is affected or not. If any affected attempts with disease are noted, the remedies are also displayed. [Sanjay et al 2013] The

implementation specifics of the work proposed are listed in this section.

Figure 1: Proposed method

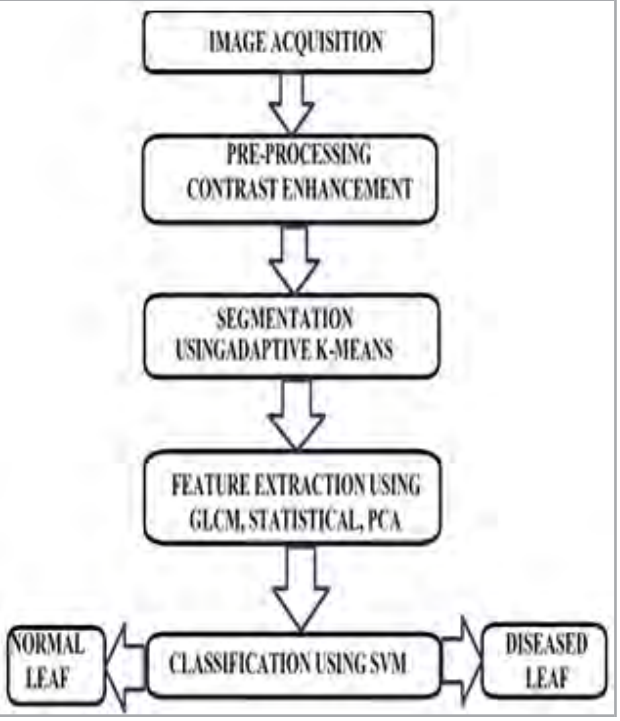
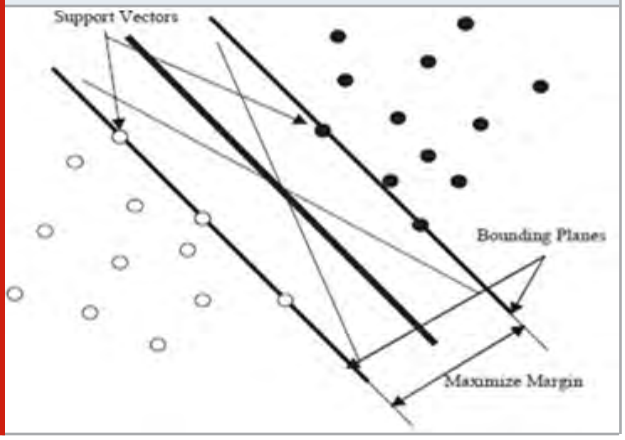


Figure 2: Bounding planes, Support vectors and Maximum Margin in SVM



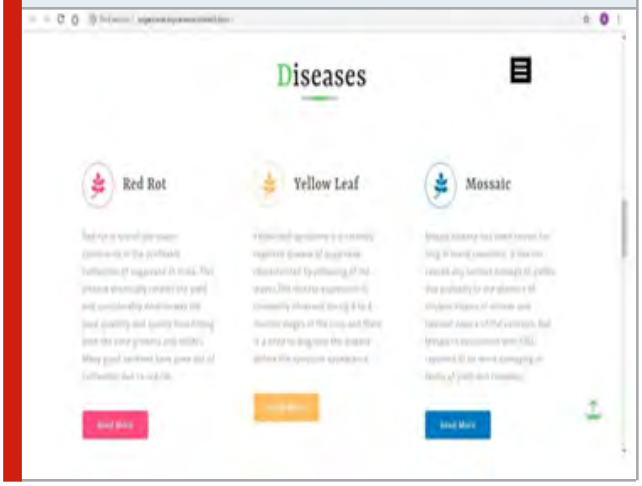
(i) Adaptive Histogram Equalization: Adaptive histogram equalization (AHE) is a computer-ized image pre-processing technique used to get a high improvement in contrasting the images. This method performs several histograms, wherein each histogram represents the distinct portion of the image, and then redistribution is done based on the image tone. Therefore, it helps in the enhancement of edges and in the improvement of local contrast of an image.

(ii) Adaptive k-Means Clustering: k-means clustering is an algorithm that comes under the unsupervised learning group. The primary intent of this k–means clustering algorithm is to construct k number of groups based on similar features. As a result, centroid is chosen for each cluster in order to create a new label data and this is followed by training the label data. This algorithm helps in identifying and analysing the groups that have evolved organically instead of segmenting groups without looking at the data. The weights of the centroid help to understand what sort of group each cluster specifies [Jagan et al.,2016]. This algorithm uses rehearsal refinement to generate the result. The inputs of the algorithm are the numbers of data set and the clusters k. The algorithms starts with the estimation of the k centroids and then repeats between three steps: Data Assignment, Centroid Update and Choosing k.

(iii) Feature Extraction: Feature extraction is a process in which the informative features extracted and it is similar to dimensionality reduction. Whenever there is redundancy in the input data or if the input data is too large, feature extraction is performed to reduce the data. In character recognition system, the feature extraction is usually done after the pre-processing phase [Reddy etal.,2017]. Feature selection is nothing but determining the subsets of the initial features. It is the critical stage in the entire process as the classifier cannot recognize from the poorly selected features. It is helpful in achieving better performance as well.

Figure 3: Home Screen of the web application

Figure 4: Leaf disease details in Web Application



(iv)Gray-Level Co Occurrence Matrix: It is one of the statistical techniques for texture analy-sis. It deals with the spatial relationship of pixels and is often referred to as the spatial dependency matrix of the gray-level. In estimating the joint probability distribution, the average occupancy level of the matrix should be greater. This is accomplished by limiting the number of levels of amplitude quantization, which could decrease the precision of the texture of low-amplitude. Each element (i,j) in the resulting GLCM represents the sum of the number of times the pixel with value i that occurred in relation to the pixel of j [Reddy et al.,2017], where i represents the intensity value and j represents the pixel value. The required step to calculate a GLCM for the complete dynamic picture is prohibitory and gray comatrix scales the input image. Gray comatrix performs scaling operation by default to reduce the intensity values of the gray scale image from 256 to 8. The number of grey levels shows the scale of the GLCM [Devi Aruna,2016].

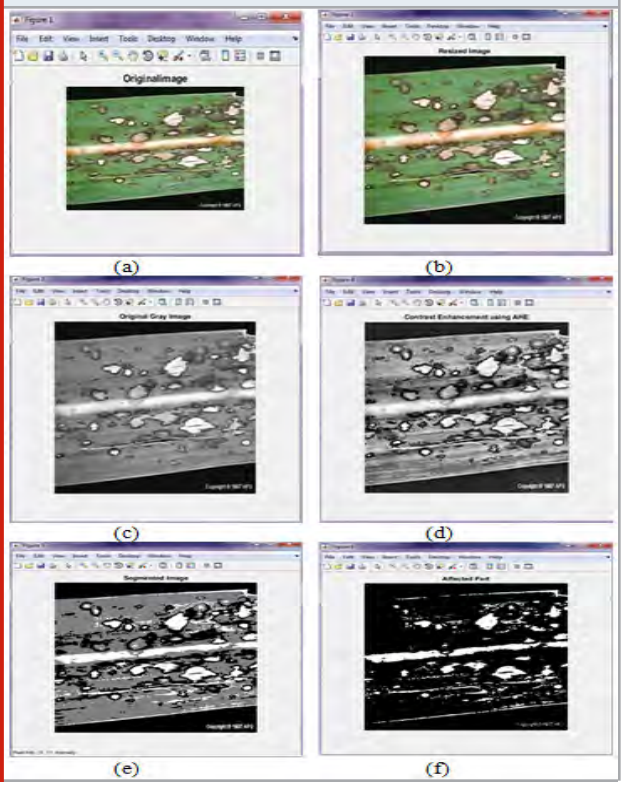
(v) Principal Component Analysis: PCA is a kind of statistical method which creates new un-correlated variables known as principal parts using orthogonal transformation. The ensuing vectors are the unrelated set of orthogonal basis. PCA is susceptible to the relative size of the initial variables. PCA is performed by eigenvalue decomposition of an information variance matrix or singular worth decomposition of an information matrix. Every attribute is standardized by mean centering such that its empirical mean value is zero. The non-standardized attributes are eigen vectors. They are the cosines of orthogonal rotation of variables into principal elements or back [B.Chitradevi, P.Srimathi,2014].

(vi)Classification: Classification deals with the wide range of decision-theoretic approaches for image identification. The image in question has one or many features and these classification algorithms helps in classifying the image to one of the distinct classes based on the features. The classes are specified in prior by the analyst [Shweta and Shandilya,2018]. The different types of well-known classifiers are Support vector machines, Quadratic classifi-ers, Linear classifiers, Learning vector quantization, Kernel estimation, Neural networks, and Decision trees.

(vii)Support Vector Machine: It is one of the well-suited algorithm for image classification. To classify the images remotely, Kernel based image classification is mostly used. SVM with multiple kernels is used for the classification of images with kernel optimization. Feature extraction is an important process that must be carried out before classification, because the images are categorized on the basis of the extracted features. SVM Kernel method provides a good solution for regression and classificationbased problems. This is a supervised machine learning algorithm and the two different classes are separated by a hyperplane. If the classes are separated with the larger margin, then it has minimal error. The minimum error is used to predict the correct class of the data without any error in the classification process [Chitradevi, P.Srimathi,2014].

There are two parallel planes called "Boundary planes" and the distance between these two planes is known as 'margin'. The points that lie on the boundary planes are called 'support vectors' [Shweta and Shandilya,2018]. It is one of the best and popular method used for image classification. The support vectors (Figure 2) are the data in the data set which defines the maximum margin. Web Application Development: The Web based application provides the control measures of identified sugarcane diseases. This application helps to identify six diseases of sugarcane and provides their symptoms and precautions. The user can access the web application using the link http://sugarcane.mycareersconnect.com/. In the application, images of infected leafs are compared using pixel by pixel comparison method. The uploaded image is matched with the images existing in the database. The hexacode of the RGB colors is calculated for each pixel of both the images and stored in two different matrices. Then these matrices are compared with each other and when the two images have more than 80% similar pixels, it would give the desired result.

Figure 5: Results for RINGSPOT: a) Input RGB image b) Resized image c) Grayscale image d) Contrast enhanced image e) Segmented image f) Classified result



The web application for our system is targeted to be used by farmers or the people who are involved with the production process of Sugarcane. It will benefit them in a way that gives them solution within a short period for the experts. Since it is a web-based application, made accessible to all the farmers regardless of their physic allocation. The home screen of web application and the leaf disease details are shown in Figures 3 and 4 respectively.

Figure 6: Results for Mossaic: a) Input RGB image b) Resized image c) Greyscale image d) Contrast enhanced image e) Segmented image f) Classified result

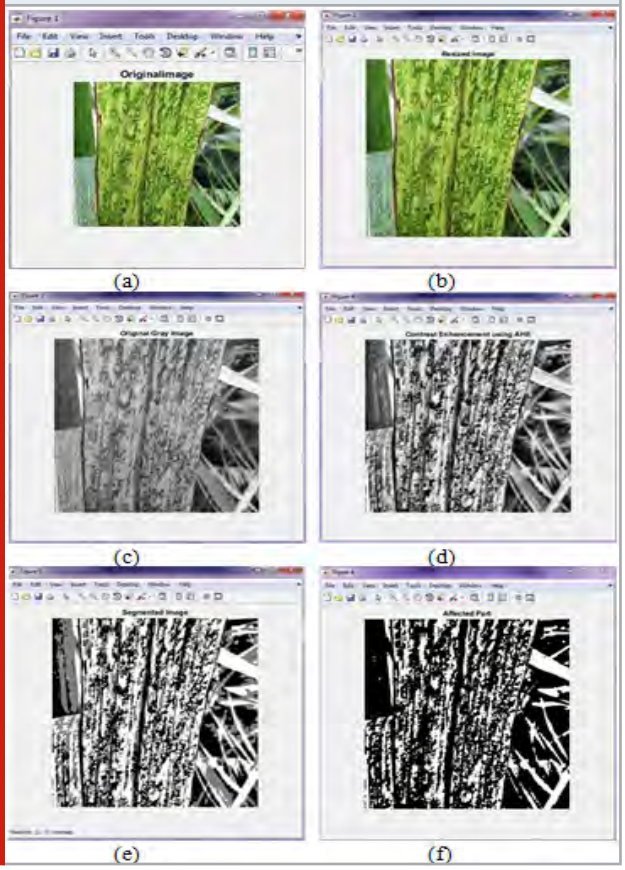


Figure 7: Results for Red Rot:a) Input RGB image b) Resized image c) Greyscale image d) Contrast Enhanced image e) Segmented image f) Classified result

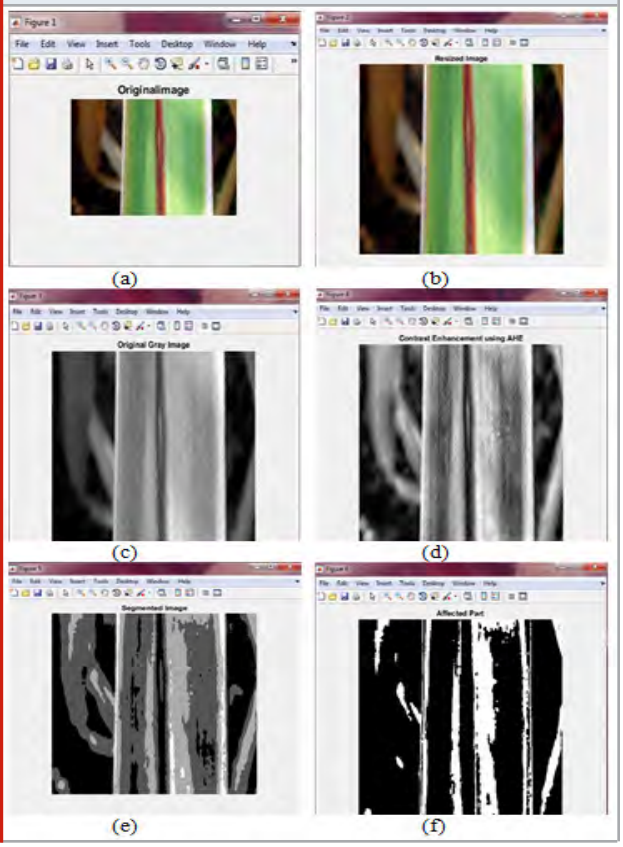


Figure 8: Results for EYESPOT: a) Input RGB image b) Resized image c) Greyscale image d) Contrast enhanced image e) Seg-mented image f) Classified result

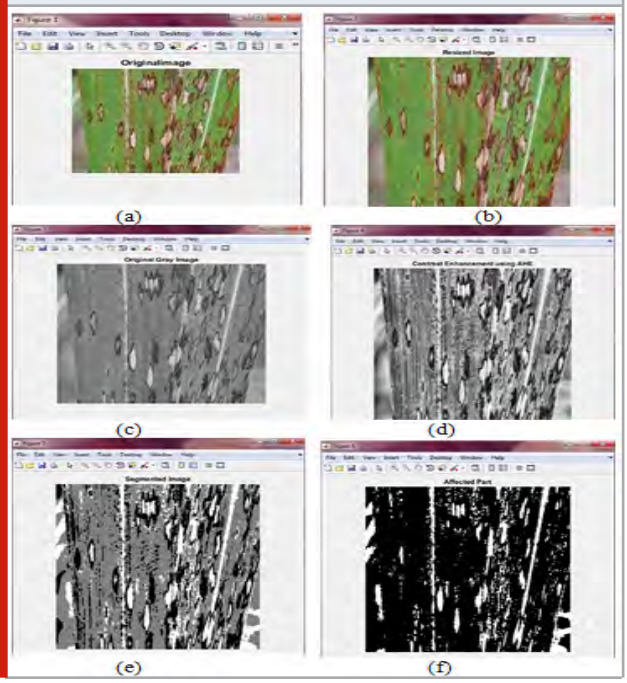
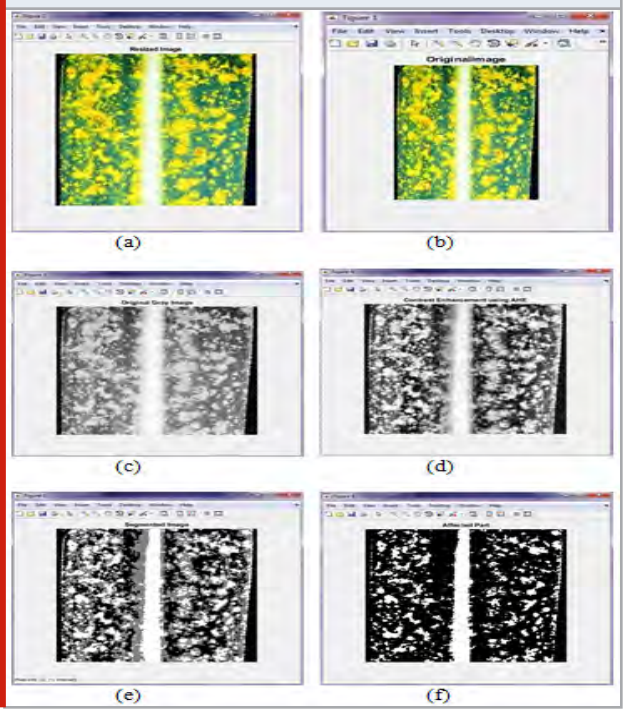


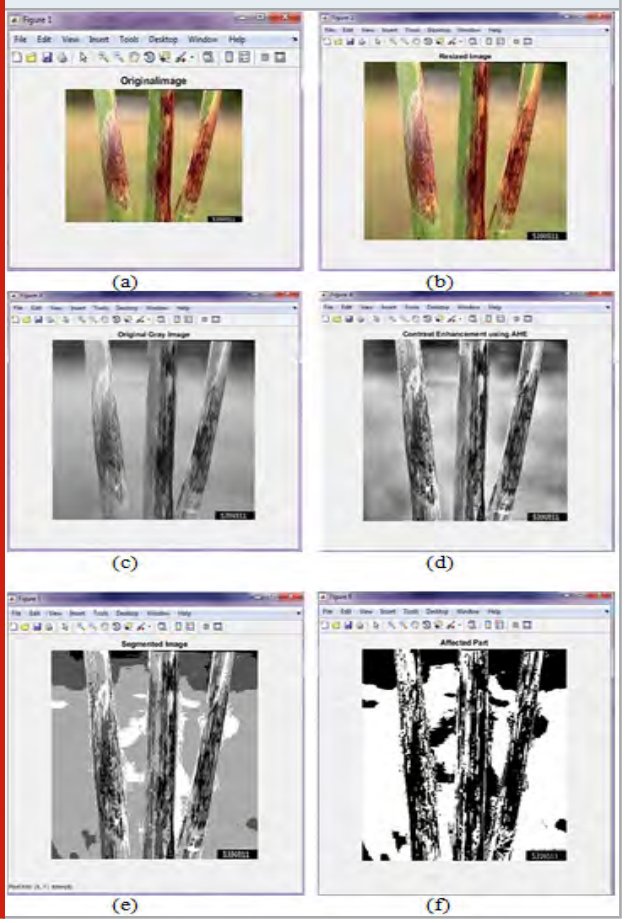
Figure. 9: Results for yellow leaf a) Input RGB image b) Resized image c) Greyscale image d) Contrast enhanced image e) Segmented image f) Classified result.



RESULTS AND DISCUSSION

The diseased leaf of sugarcane is acquired and through image pre-processing technique where the unwanted image data are suppressed. The image is enhanced through AHE. Then, the process of segmentation is done through Adaptive k-Means Clustering. The desired features are extracted by using GLCM and PCA. Finally, the classification of diseases is done through SVM. The above described process is executed in MATLAB software and the obtained results for six sugarcane diseases are shown in figures from 5 to 10.

Figure 10: Results for leafscald: a) Input RGB image b) Resized image c) Greyscale image d) Contrast enhanced image e) Segmented image f) Classified result



CONCLUSION

In present situation, it is difficult for farmers to keep an eye on each plant in the growing area and detect the manifestation of any infection. To serve the function of a watchdog by detecting the disease and determining the form of the disease in the leaf, image processing techniques have been developed. The combination of various feature extractions (GLCM and PCA) with SVM classifier has been implemented to test six signifi-cant diseases which largely affect the sugarcane yield. The proposed system achieved the ac-curacy value of 95%. A detailed review of the causes and symptoms of all sugarcane diseases is highlighted in this article. The web application developed for our system is targeted for the farmers to apply it. It will benefit them with desirable solutions within a short span of time. As a future work, various classification methods can be tried to increase the efficien-cy of the system in which and the system can be improved by a broader data set for the identi-fication of all forms of sugarcane diseases. For the web application part, the information can be made available in the regional language.

ACKNOWLEDGMENTS

This work was funded by Tamil Nadu State Council for Science and Technology under Stu-dent Project Scheme. Authors would like to thank ICAR-Sugarcane Breeding Institute, Coim-batore for providing the database of sugarcane leaf disease to carry out the work.

Conflicts of interest: The authors declare no conflict of interest.

REFERENCES

Al-Hiary H, Bani-Ahmad S, Reyalat M, Braik M and ALRahamneh Z 2011, "Fast and accu-rate detection and classification of plant diseases", International Journal for computer applica-tions, Vol 17 No 1. Pages 31-38.

Baddeli sravya reddy, Deepa R, Shalini S, Bhagyadivya P 2017, "A novel machine learning based approach for detection and classification of sugarcane plant disease by using DWT", International Research Journal of Engineering and Technology, Vol 04 No 12.Pages 843-846.

Chitradevi,B, P.Srimathi 2014, "An overview on image processing techniques". Interna-tional Journal of Innovative Research in Computer and Communication Engineering, Vol.2,No 11. Pages 6466-6472.

Devi Aruna.D 2016,"A Survey on Different Disease and Image Processing Techniques in Sugarcane Crops",IJSRD International Journal for scientific Research & Development, Vol.3 No 11.Pages 323-325.

Dheeb AlBashish, Malik Braik and Sulieman Bani-Ahmad 2011,"A framework for detection and classification of plant leaf and stem diseases", 2010 International Conference on Signal and Image Processing. 10.1109/ICSIP.2010.5697452.Pages 113-118.

https://plantvillage.psu.edu/topics/sugarcane/infos

https://sugarcane.icar.gov.in/index.php/en/2014-04-28-11-31-50/major-diseases

Jagan Bihari Padhy, Devarsiti Dillip Kumar, Ladi Manish and Lavanya Choudhry 2016,"Leaf Disease Detection Using k Means Clustering And Fuzzy Logic Classifier", In-ternational Journal of Engineering Studies and Technical Approach, Vol 02, No. 5.Pages 1-7.

Sandesh Raut, Amit Fulsunge 2017, "Plant Disease

Detection in Image Processing Using MATLAB". Vol 6 No 6. Pages 10373-10381.

Sanjay B. Dhaygude, Nitin P.Kumbhar 2013,"Agricultural plant Leaf Disease Detection Using Image Processing International Journal of Advanced Research in Electrical, Elec-tronics and Instrumentation Engineering Vol. 2 No 1.Pages.599-602.

Shweta R. Astonkar, Shandilya VK 2018, "Detection and Analysis of Plant Diseases Using Image Processing Technique International Research Journal of Engineering and Technology Vol.05 No.04.Pages 3190-3193.

Vijaisingh, Misra AK 2017, "Detection of plant leaf diseases using image segmenta-tion and soft computing techniques", Information Processing in Agriculture, Vol 4 No 1 Pages 41-49.

Viswanathan R, Rao GP 2011," Disease Scenario and Management of Major Sugarcane Diseases in India", Sugar Tech, Vol 13 No 4, Pages 336-353.

Biosc.Biotech.Res.Comm. Special Issue Vol 13 No 11 (2020) Pp-116-120

BBRC
Bioscience Biotechnology Research Communications

Classification of Leucocytes Using Deep Learning

Suganthi. N[1], Preethi. V[2], Swetha. K[3] and Kannan. K[4]
[1]*CSE department, Kumaraguru College of Technology, Coimbatore, Tamilnadu, India*
[2,3,4]*Information Technology, Kumaraguru College of Technology, Coimbatore, Tamilnadu, India*

ABSTRACT

This work is about categorization of white blood cells (WBC) or leucocytes has become highly crucial for the analysis of anaemia, leukaemia and many other hematologic diseases. Out immune system is dependent on the WBC concentration in our blood. The aim of this paper is towards developing a computerized WBC category system using deep learning. Many models proposed for this application so far has used transfer learning by fine tuning the ResNet, Inception and VGGNet. But all these models were trained on ImageNet dataset which is completely different from the dataset used in this application. So, we have proposed a deep learning model for the white blood cells classification task from the scratch, without using transfer learning.

KEY WORDS: CONVOLUTION NEURAL NETWORK, CONVOLUTION LAYER, LEUCOCYTES, POOLING LAYER, WHITE BLOOD CELLS.

INTRODUCTION

In recently, hematology analysis has been a significant role in diagnosing many fatal diseases. There are many causes for an elevated leucocyte count like overproduction, or body issuing WBC early from the bone core. Sometime in worst case, new WBC, so-called blasts, accumulate in infected area. Some reasons of an enhanced leucocytes amount include stress, serious infections, bone marrow injury or illnesses like aplastic anemia, bone marrow. There are several kinds of leucocytes in our blood. In the blood test, the different proportion of these different leucocytes indicate several health conditions. Mostly the ratio of these leukocytes remains steady in blood, any disorder of percentage of WBC is a very consistent evidence to indicate that the blood sample is drawn from an ill patient.

ARTICLE INFORMATION

*Corresponding Author: suganthi.n.it@kct.ac.in
Received 20th Oct 2020 Accepted after revision 12th Dec 2020
Print ISSN: 0974-6455 Online ISSN: 2321-4007 CODEN: BBRCBA

Thomson Reuters ISI Web of Science Clarivate Analytics USA and Crossref Indexed Journal

Clarivate Analytics crossref

NAAS Journal Score 2020 (4.31) SJIF: 2020 (7.728)

*Online Contents Available at: http//www.bbrc.in/
Doi: http://dx.doi.org/10.21786/bbrc/13.11/25*

MATERIAL AND METHODS

American and Japanese companies created WBC identification methods in 1970's. The image recognition issue of WBC was examined. It is used owing to more price and minimal correctness (Mazin Z et al 2017). A system (Yang X et al 1994) founded on mathematical morphology (TSMM) were created. The authors (Ushizima D M et al 2005) studied the option of using SVMs in identifying various kinds of leukocytes. Another proposal (Gaobo Liang et al 2018) suggested a novel discovery process built on cellular neural networks. This work was mostly assessed with a reduced amount of data, thus lacked good generalization capability.

A framework (Shubham Manik 2016) is proposed to increase detection and classification of leucocytes using image segmentation and nucleus improvement techniques to extract cell features. The above is fed into artificial neural networks for estimate(Alex Krizhevsky2010) Recently, an automated leucocytes classifier method has been proposed which uses transfer learning using various state of the art models and use voting module to predict the class with maximum score in all the models (Wei Yu 2017). Another proposed prototype WBCNet (Ming Jiang 2018) that can get aspects of the microscopic WBC

picture by merging batch normalization process, residual convolution design and enhanced activation function.

The dataset is taken from kaggle datasets (Paul Mooney 2018). The initial data set contains of 352 pictures (RGB) of WBC of resolution 640x480. It is composed of 3 Basophils, 88 Eosinophils, 33 Lymphocytes, 21 Monocytes and 207 Neutrophils shown in figure 2. Since the count of basophils is very limited in our dataset, we have ignored it to propose a generalizing model. Since the dataset is imbalanced with varying number of samples for each class, data augmentation techniques are used. Thus, the final dataset is 12444 pictures composed of 3120 Eosinophils, 3103 Lymphocytes, 3098 Monocytes, and 3123 Neutrophils. These augmented images are of resolution 320x240. The sample augmented images used in training are shown in figure.1.

Figure 1: Microscopic image samples of four types of WBCs

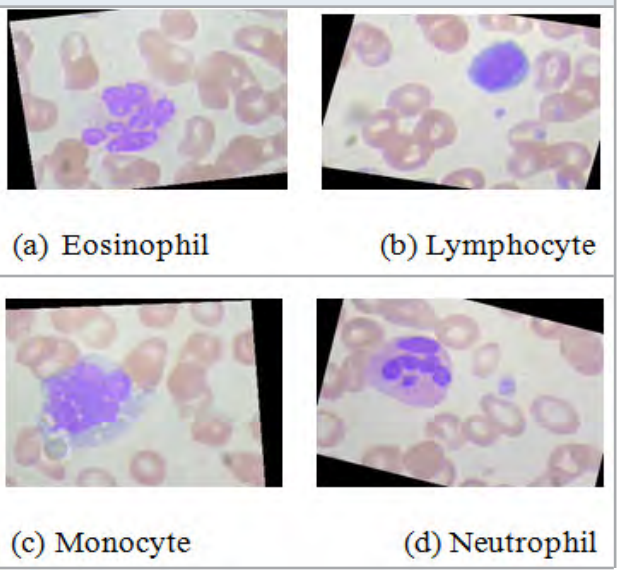
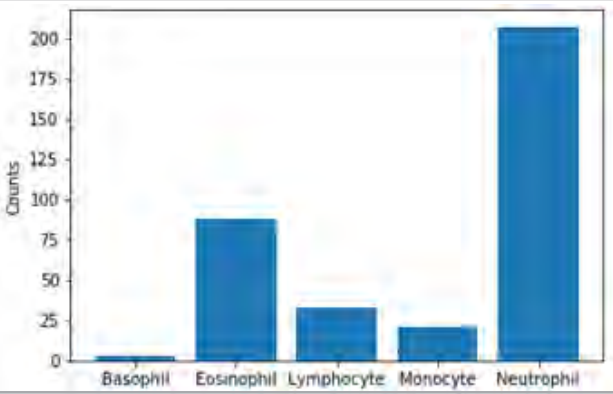


Figure 2: Data distribution



The dataset is comprised of two folders –train (9957 samples) and test (2487 samples). The train data is split into two for training (80%) and validation (20%). Training data consists of 7965 samples with 1998 Eosinophils, 1986 Lymphocytes, 1982 Monocytes and 1999 Neutrophils. Validation data consists of 499 Eosinophils, 497 Lymphocytes, 496 Monocytes and 500 Neutrophils. The test data consists of 623 Eosinophils, 620 Lymphocytes, 620 Monocytes and 624 Neutrophils. The model does not require any intensive preprocessing techniques. The images are reshaped to resolution 120x120 with three channels. Then they are feature normalized by dividing each pixel value by 255 for all channels. The labels are encoded and then converted to categorical one-hot labels which are used throughout the model.

CNN most popular deep learning architectures. CNN design is highly invariant to translation, tilting, scaling etc. It also avoids the difficult preprocessing of images prior to training and it is and end to end process. Because of these reasons, CNN has highly gained the attention. There are three main characteristics of CNN are local convolution, weight sharing (same weights learnt are shared to detect features throughout the image) and multi-kernels of convolution. The different kernels in CNN help detect different features. One of the most important aspects of CNN is that a convolution layer needs to learn relatively very small number of weights compared to that learnt in a conventional fully connected feed forward neural network. This requires comparatively less memory to save the weights. Thus, the features learnt in convolution layer are fed further to traditional fully connected layers to generate output. The outcome shows the possibility that a picture be appropriate to one of multiple classes using activations like sigmoid (binary classification) or softmax (multi-class classification).

The design of this model is summarized in the Figure.3. It consists of six learned layers – three convolution layers and three connected layers. It also consists of three pooling layers stacked between convolution layers. The convolution layers use very small receptive fields of 3x3 kernels instead of using large receptive fields. These small kernels stacked up give better results compared to models using single large receptive field. Thus, we have used 3x3 kernels for all convolution layers in our model. The weights to be learnt in the model are less in convolution layers compared to those in dense layers.

We have adapted an interesting idea for stacking up convolution layers from VGG16 network (Karen Simonyan 2015). The convolution layers are stacked up in such a way that the dimension of image reduces as it passes through multiple layers and the number of channels increase with increase in number of filters used in the convolution layers. This helps model to learn more complex features in the later layers. To prevent convolution layers from overfitting, we have used small regularizers in second and third convolution layers. Throughout the model, we have used max-pooling layers with kernel size 2x2 and stride 2 to reduce the image size by half as it passes the convolution layers. This layer also allowed model to ensure the features predicted in convolution layers by taking maximum values for each pool area (2x2) across the image.

Most important features to be extracted by the convolution layers are the cell boundary and the shape of the nucleus. Thus, we have used 'same' padding in all convolution layers to ensure that the boundary information is not lost as the image passes through the layers of the model. Relu non-linearity has been used in all convolution layers and two dense layers. This activation helped the model to avoid oscillating in undesirable local optima or vanishing gradients problem.

$f(x) = \max(0,x)$

The last dense layer uses softmax activation to classify the leucocytes out of four different classes. The output of softmax is the probability of prediction for four classes and the predicted class is the one with maximum probability.

$$\sigma(x_j) = \frac{e^{x_j}}{\sum_i e^{x_i}}$$

The three dense layers are stacked up in such a way that the number of hidden units in each layer decrease in order. 97% of the network parameters are in dense layers. This over fitted the training data and reduced the generalization in validation and testing. Thus, we have used dropout technique in dense layers to allow them to learn less dependent on the incoming activations.

RESULTS AND DISCUSSION

We conducted this experiment in Google colab jupyter notebook on GPU(K80). We coded the model using keras library with tensor flow as backend. We trained different models with slight changes in architecture and got different results. During training, we used Data augmentation technique using keras Image Data Generator which applies simple transformations on images like shifting and flipping. We trained images on batches with batch size-32 in all models. 'Xavier uniform initializer' is used for initializing the weights in convolution and dense layers. Increasing or decreasing the learning rate did not perform well even in training. We shuffled the data randomly before training and before each epoch which prevented the model from memorizing the predictions based on the order of images trained. The evolution of the proposed model (model-6) is given in the figure 4 and 5.Images with different dimensions (160x120) where used in model-1, while the other models where trained on 120x120 images to ensure uniform padding in all convolution layers. But reducing image size reduced the accuracy of model-2 in both training and testing. Model-1 and Model-2 used 'valid' convolution which reduced image size across the convolution layers which is undesirable for this task. So, we used 'same' convolution to retain the image size in all layers for the rest of the models. Model-3 gave better accuracy using 'same' convolution.

Model-4 is the largest model which uses third convolution layer with 128 filters followed by dense layer with 128 activation units. This model outperformed the other models on training but did not generalize well on testing which clearly shows that the number of parameters learnt over fit the available training data and did not perform well on unseen data. So, we applied batch normalization for convolution layer with 128 filters and dense layer with 128 activation units. This improved accuracy of the model but did not perform well compared to previous model. This model also produced lot of fluctuations while training and created many unwanted peaks in accuracy. We also applied decay for learning rate but it did not help to improve the accuracy after few epochs.

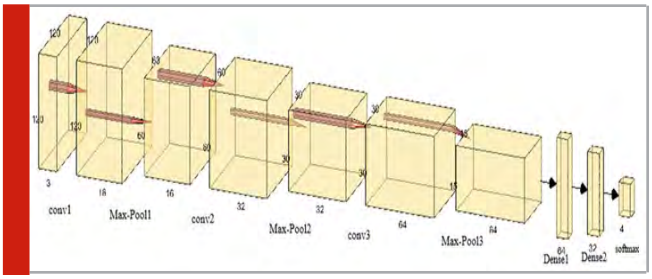


Figure 4: Loss Curves

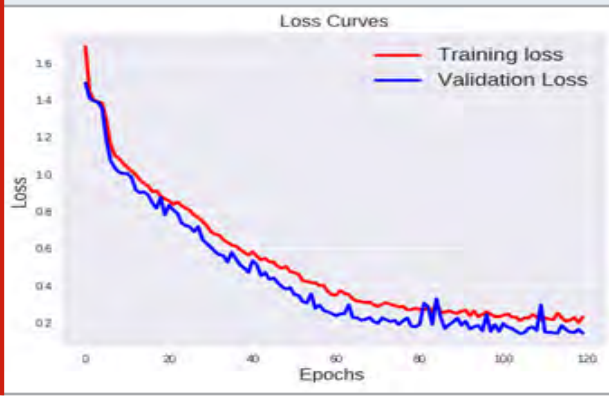
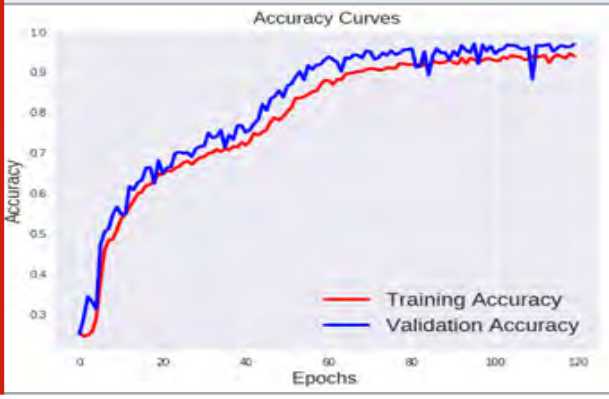


Figure 5: Accuracy Curves



The model was oscillating in local optima for long time without any improvement in accuracy. Model-5 which is close to our final model has less parameters compared to model-4. It uses regularization in a convolution layer and dropouts in dense layer. This model performed like model-4 on training and it also gave good accuracy in

validation. But it gave very less accuracy on testing. Though validation data is unseen data for the model, it over fit the validation data and so it failed to perform well on test data. Model-6 which is our final model gives very good accuracy in training and generalizes well on validation and testing. It uses dropouts like model-5 and uses regularization in two convolution layers with regularization constant-0.01.

Figure 6: Comparison between different model accuracies

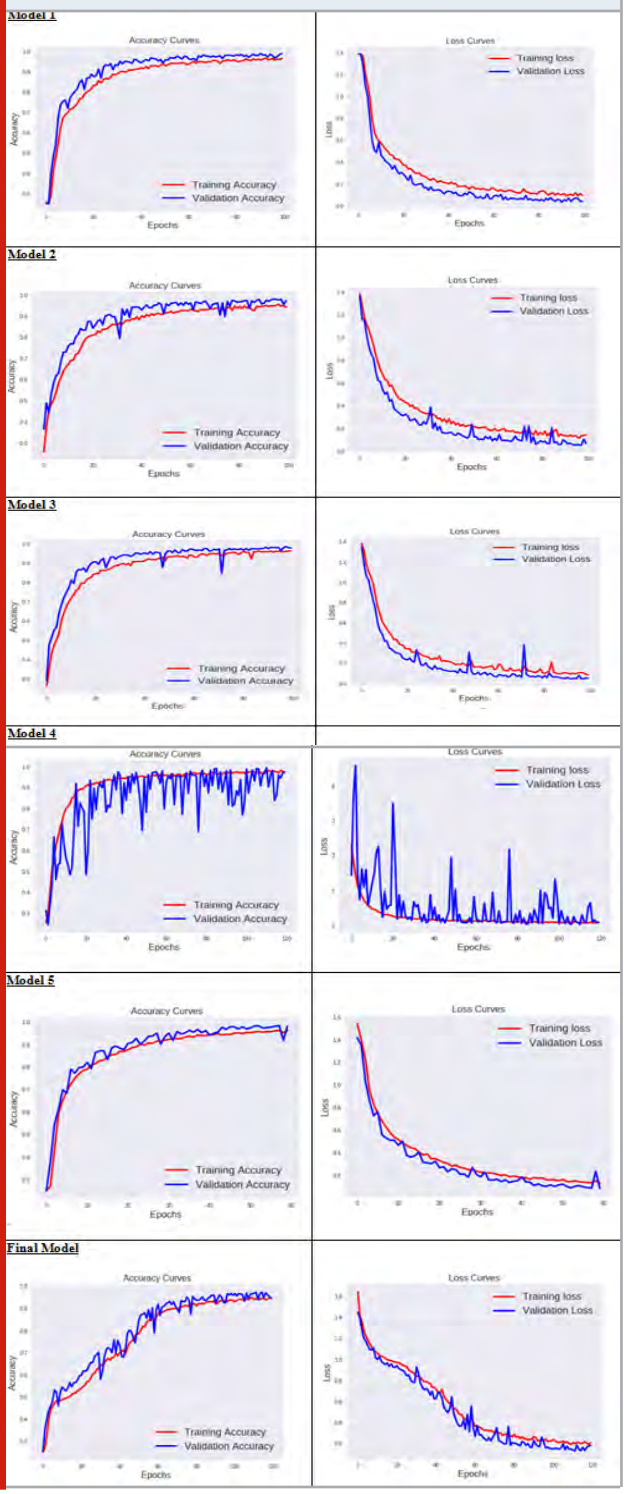


Figure 7: Confusion Matrix

	Eosino phils	Lympho cytes	Mono cytes	Neutro phils
Eosinophils	531	14	0	78
Lymphocytes	0	620	0	0
Monocytes	0	0	580	40
Neutrophils	53	1	12	558

The metrics measured are accuracy (Top_1 and Top_2) and loss. The graphs above show the loss and accuracy obtained during training and validation of our final model. The model gives accuracy of 97.0% on training and 96.7% on validation. The model generalizes very well on test set with correctness rate of 92.0%. The confusion matrix for prediction on the test dataset is given below in figure.7. Our model performs with 100% accuracy in predicting Lymphocytes and Monocytes with accuracy of 93.5%. This is followed by Neutrophils 89.4% and Eosinophils (85.2%). All of the misclassified monocytes and most of the eosinophils are confused by the model to be neutrophils. The model is trained for 120 epochs. Further training the model makes it to memorize the training data and reduces the generalizing capability of the model.

CONCLUSION

The proposed model can be further improved by training on more data and using large models. Mode-6 suffered bias when trying to regularize overfitting (bias –variance trade off). This can be avoided if we train very deep models. But large volume of data is required to develop such models. Thus in future, more accurate model can be proposed to avoid labor intensive manual white blood cell classification.

REFERENCES

Alex Krizhevsky ,Sutskever, Geoffrey E. Hinton (2010) " ImageNet Classification with Deep Convolutional Neural Networks(AlexNet)", ImageNet LSVRC-2010 contest, NIPS'12 Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1,Pages 1097-1105

Gaobo Liang et al (2018) "Combining Convolutional Neural Network With Recursive Neural Network for Blood Cell Image Classification", Special Section On Trends, Perspectives And Prospects Of Machine Learning Applied To Biomedical Systems In Internet Of Medical Things, Vol – 6, page: 36188 – 36197

Karen Simonyan_ & Andrew Zisserman (2015) "Very deep convolutional networks for large-scale image recognition (VGGNet)", ICLR-2015, arXiv:1409.1556

Mazin Z. Othman,Thabit S. Mohammed,Alaa B. Ali (2017) " Neural Network Classification of White

Blood Cell using Microscopic Images", (IJACSA) International Journal of Advanced Computer Science and Applications,Vol.8,Issue.5

Ming Jiang et al (2018) White Blood Cells Classification with Deep Convolutional Neural Networks, International Journal of Pattern Recognition and Artificial Intelligence, Vol. 32, No. 09, 1857006

Paul Mooney (2018) Blood Cell Images dataset-12,500 images: 4 different cell types,version-6

Shubham Manik;Lalit Mohan Saini and Nikhil Vadera (2016) "Counting and Classification of White Blood Cell using Artificial Neural Network (ANN)", 1st IEEE International Conference on Power Electronics. Intelligent Control and Energy Systems ICPEICES-2016), doi-10.1109 /ICPEICES. 2016.7853644

Ushizima D.M, A. C. Lorena, and A. de Carvalho (2005) "Support vector machines applied to white blood cell recognition," in International Conference on Hybrid Intelligent Systems, 2005, p. 6 pp.

Wei Yu et al (2017) "Automatic Classification of Leukocytes Using Deep Neural Network", 2017 IEEE, 12th International Conference on ASIC (ASICON), pp. 1041-1044

Yang .X , L. Luo, and Y. Wei (1994) "Automatic classification system for leukocytes in human blood," Chinese Journal of Computers, vol. 17, no. 2, pp. 130–136.

Biosc.Biotech.Res.Comm. Special Issue Vol 13 No 11 (2020) Pp-121-125

BBRC
Bioscience Biotechnology Research Communications

A Novel Hybrid Method for Classification of Tumor in Gene Expression Based Central Nervous System Microarray Data

W. Jai Singh[1]* and R. K. Kavitha[2]

[1&2]*Assistant Professor (SRG), Department of Computer Applications*
Kumaraguru College of Technology, Coimbatore, India

ABSTRACT

Nowadays, DNA microarray is widely used by researchers to predict cancer disease. In a microarray data, the presence of large number of features and instances makes it difficult to analyse and diagnose cancer. Hence, the selection of features is considered to be a vital task in classifying data. The phenotypical and genotypical behaviour of tumors in the human central nervous system poses a challenge in diagnosing and treating the disease. A clear description of the tumor is necessary to diagnose and treat the disease. Many feature selection methods can be used to identify the genes which are expressed differently in a microarray data. This research proposes a novel method of categorizing tumors present in Central Nervous System (CNS) with the help of DNA microarray gene representation present in the samples collected from patients. This research work aims to blend techniques like Support Vector Machine (SVM), Information Gain (IG) and Principal Component Analysis (PCA). Initially, Information Gain technique was used to select features followed by feature reduction with the help of principal component analysis. Finally, support vector machine was employed to classify the disease as cancer or not. This study was carried out with the popularly used CNS microarray dataset. It can be observed from the study results that the proposed approach proves to be effective in yielding a superior accuracy in classifying the disease with minimal number of genes.

KEY WORDS: CLASSIFICATION, DIMENSIONALITY REDUCTION, FEATURE SELECTION, MICROARRAY, SUPPORT VECTOR MACHINE.

INTRODUCTION

The tumor found in the human central nervous system can be an unusual cell growth in brain or the spinal cord tissues. CNS tumor which is a universal term comprises of one hundred and twenty individual types of tumors. It has been discussed in few studies that the identification and forecast of these tumors post a great challenge due its varied behaviour (NCI 2005, Yang et al 2003). The study of Gene expression using microarrays is an important area in the field of biomedical engineering. Latest improvements in DNA microarray technology expedited analysing and assessing the representation levels of a thousands of genes at identical time, so permitting a large generation of microarray data (Fan et al 2009).

Microarray techniques have been effectively utilized widely in the field of biomedical research since they provide the likelihood of performing huge number of tests on genome patterns (Vilda et al 2006). The gene expressions of the microarray consist of several dimensions. The evaluation of each gene is made possible in a particular environment where various types of cancerous tissues are found (Zheng et al 2006). The classification process of microarray is carried out

ARTICLE INFORMATION

*Corresponding Author: jaisingh.w.mca@kct.ac.in
Received 15th Oct 2020 Accepted after revision 12th Dec 2020
Print ISSN: 0974-6455 Online ISSN: 2321-4007 CODEN: BBRCBA

Thomson Reuters ISI Web of Science Clarivate Analytics USA and Crossref Indexed Journal

NAAS Journal Score 2020 (4.31) SJIF: 2020 (7.728)

Online Contents Available at: http//www.bbrc.in/
Doi: http://dx.doi.org/10.21786/bbrc/13.11/26

in two stages. The initial stage is to choose a group of important and appropriate genes and the next stage is to construct a classification model with which data can be predicated with a higher accuracy. For a precise diagnosis and curing of the cancer disease, it is important to have a good classification model. The presence of immense dimensions in a DNA microarray data poses a problem when used for classifying cancer disease. The reason for this can be the small sample size of DNA micro array when compared to gene size (Du et al 2014).

Normally, a small portion of the genes will be effective for classifying data in a large gene dataset. Hence choosing the appropriate genes plays a significant role in studying microarray data. Also, the correct choice of genes leads to accurate classification of the samples. In case of supervised classification technique, a high-level of precision in data classification can be achieved by selecting appropriate features. Moreover, the technique offers an improved classification quality and reduced computational complexity of the chosen algorithm. The dimensionality of the classification technique can be decreased by obtaining the smallest number of features out of the actual features thus leading to an accurate classification of data. By performing this, inappropriate and unneeded features can be eliminated.

The proposed work challenges the classification of cancer disease with the help of gene expression summaries. An innovative method of examining the microarray dataset to classify cancer is presented in this paper. The novel method initially used Information Gain for selecting features and then uses PCA for reduction of features and ultimately uses SVM for classifying cancer disease. The proposed method enhances the precision of cancer classification by lowering tohe feature count. The performance of the recommended methodology is assessed by applying it on CNS cancer dataset. The classification outcomes are matched with the findings obtained from other recent approaches.

MATERIAL AND METHODS

Literature Review: The cancer disease can be diagnosed by detecting it at an early stage with the help of a new technique called microarray gene expression. The extremely important genes which may be the root cause of disease can be detected easily by classifying the gene expression data. The classified data helps in treating the cancer patients in an effective way. Feature selection helps in dimensionality reduction and redundancy elimination of the data in an efficient manner during the process of classification. Tan et al. (2008) presented a feature selection framework by combining GA and other methods of feature selection. It was concluded that hybrid approaches prove to be more useful and stronger when compared to applying individual algorithms or techniques on the dataset.

Alexander Statnikov et al (2005) has developed a reliable cancer investigative style centred on microarray data. To offer the model with the ideal blend of classifier, the researcher has used cross validation and genre methods. Also, a detailed evaluation of many algorithms was performed. A novel mixture of feature selection or extraction with Naïve Bayes was used by Rabia Aziz et al (2015) to classify the cancer in huge number of attributes in microarray data. Some of the pre-processing methods like ICA and filtering technique used for extraction of an attributes. Luque-Baena et al.,(2013) have used two approaches namely genetic algorithm and stepwise forward selection for analysis of microarray data. The genetic algorithm approach mixes the mutual information and classification techniques to foresee the cancer disease in patients.

Ding and Peng (2005) has recommended a minimum redundancy maximum relevance (MRMR) feature selection framework to get rid of redundancies observed in microarray gene expression data. Huang and Chow (2005) presented a valuable feature selection proposal by assessing the mutual information built on a supervised data compression algorithm. Zhang et al. (2009) utilized mutual information for multi-label classification and confirmed that this technique enhanced the performance of multilabel classifiers.

Proposed Approach: The nearness of numerous inconsequential and insignificant highlights debases the quality of the investigations of illnesses like cancer. To counter this, it is fundamental to dissect the dataset from the correct point of view. This segment presents an approach for classifying microarray information, which comprises of two stages: i) Feature Selection and Reduction and ii) Classification. The suggested strategy is the combination of IG, PCA and SVM classifier. The point by point portrayal of the proposed strategy is as takes after:

Feature Selection Method using Information Gain: The proposed procedure utilizes IG to select the basic highlights (incorporate assurance (FS)) from the input plans of the quality microarray dataset. Information Choose up (IG) is an entropy-based highlight assessment strategy, broadly utilized inside the field of machine learning. As Information Choose up is utilized in highlight assurance, it is characterized as the whole of information given by the highlight things for the substance category. Information choose up is calculated by how much of a term can be utilized for classification of information, to degree the importance of lexical things for the classification.

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

In the above equation, pi pi is the nonzero likelihood that an self-assertive tuple in Dataset D belongs to class C_i and is estimated by $|c_i,D|/|D|$.

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j).$$

The term |Dj|/|D | acts as the strength of the jth segment. Data pick up is characterized as the distinction between the first data prerequisite (i.e., based on fair the extent of classes) and the modern prerequisite (i.e., obtained after dividing on A). That is,

$$Gain(A) = Info(D) - Info_A(D)$$

Choose the highest Gain value and add it to selection set, and repeat this process until the needed attributes are selected.

Attribute Reduction using Principle Component Analysis (PCA): PCA looks for k n-dimensional occasions that can best be utilized to speak to the information, where k <= n. The initial information is hence anticipated onto a much littler space, coming about in dimensionality lessening. PCA points to discover the headings of most extreme change in high-dimensional information and ventures it onto an unused subspace with equal or fewer measurements than the initial one. The calculation for PCA are as takes after:

1. Normalize the N number of attributes in microarray CNS dataset.
2. Compute the covariance matrix from normalized dataset
3. Find out the eigenvectors (foremost component) and eigenvalues by decomposing the covariance matrix.
4. Sort the eigenvalues by eigenvectors.
5. Select k eigenvectors which compare to the k biggest eigenvalues, where k is the dimensionality of the modern include subspace (k ≤ d).
6. Build a projection framework W from the largest k eigenvectors.
7. Change the N attribute input CNS dataset X utilizing the projection framework W to get the new k-attribute include subspace.

Classification using Support Vector Machine: The SVM may be a state-of-the-art classification strategy presented in 1992 by Boser. The hypothesis of SVM is based on the thought of auxiliary risk minimization. The microarray data set has been classified as a two class problem where the two classes are cancer and normal. To begin with, let vector $x \in R^n$ indicate a design to be classified, and let scalar y signify its category name (i.e., $y \in \{\pm 1\}$).

Additionally, let $\{(x_i, y_i), i = 1,2,...,n\}$ indicates the set of n number of samples for training, where every record in the dataset has a known class labelled as . The matter is to work out a classifier f(x) (i.e., a choice work) that will appropriately classify an input design. After applying the Information Gain and PCA from the CNS dataset, the classification technique called SVM is applied to classify the given dataset into cancer and normal.

RESULTS AND DISCUSSION

The comes about gotten utilizing our proposed calculation on the CNS microarray dataset is examined in a successive way. The CNS dataset presents the resulting forecast of the patients for embryonal tumors. This contains an add up to of 60 tests (21 are survivors and 39 are disappointments) with 7129 number of genes. This is often utilized for 2-class classification issue. A combination of techniques like information Gain, Principal Component Analysis and SVM were used to classify data and also a comparison was drawn with the results obtained with K-nearest-neighbors (KNN) and Decision Tree Algorithms.

Table 1. Classification Results using CNS dataset with all features

Classifier	Classification Accuracy	AUC	Precision	Recall
kNN	0.600	0.577	0.563	0.600
Decision Tree	0.567	0.579	0.617	0.567
SVM	0.650	0.217	0.423	0.650

Table 2. Classification Results using CNS dataset with selected features – Information Gain

Classifier	Classification Accuracy	AUC	Precision	Recall
kNN	0.783	0.875	0.793	0.783
Decision Tree	0.800	0.752	0.797	0.800
SVM	0.833	0.875	0.838	0.833

Table 3. Classification Results using CNS dataset with PCA

Classifier	Classification Accuracy	AUC	Precision	Recall
kNN	0.583	0.558	0.521	0.583
Decision Tree	0.617	0.512	0.596	0.617
SVM	0.633	0.529	0.419	0.633

Table 4. Classification Results using CNS dataset with combination Information Gain and PCA

Classifier	Classification Accuracy	AUC	Precision	Recall
kNN	0.783	0.875	0.793	0.783
Decision Tree	0.617	0.563	0.604	0.617
SVM	0.878	0.896	0.870	0.873

The performance of the classifier was evaluated using classification accuracy, precision, Recall, ROC curve and Area Under the Curve (AUC). To illustrate the adequacy and practicality of the proposed strategy, the comes about of the other classification strategies are too displayed in Tables 1 to 4.

In Table 1, the CNS microarray data are classified by SVM, Decision Tree and KNN right away with all the features. The features selected by Information Gain and PCA was applied in the data set. The Table 2 highlight the experimental results. In Table 4, the combination of Information Gain, PCA and different classification techniques are applied in the dataset. It can be seen from Tables 1–4, the IG+ PCA+SVM perform better than other individual and combinational methods, which demonstrates the effectiveness of the proposed approach. It is evident from the proposed method that the classification accuracy is high with SVM classifiers. Accurate results with reduced variation of classification performance was obtained when compared to other gene selection techniques. Thus, the proposed method displays improved classification of CNS microarray data. The performance comparison of several other methods with the proposed method using ROC Curve can be inferred from Fig. 1–4.

Figure 1: AUC curve - Classification Results using CNS dataset with all features

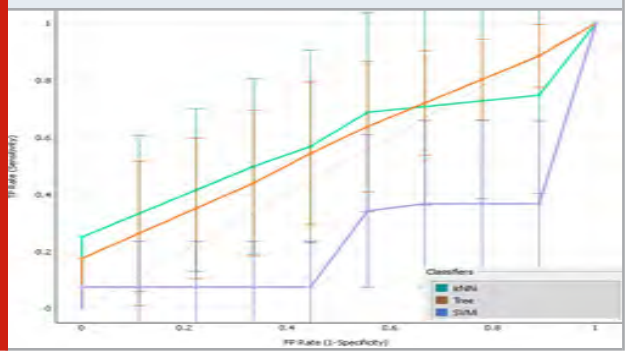


Figure 2: AUC Curve - Classification Results using CNS dataset with selected features – Information Gain

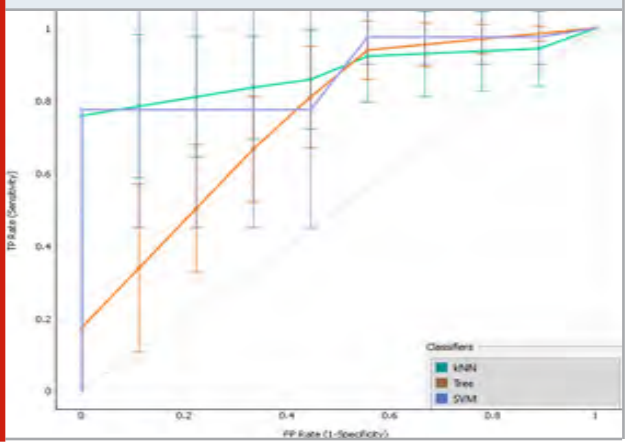


Figure 3: AUC Curve - Classification Results using CNS dataset with PCA

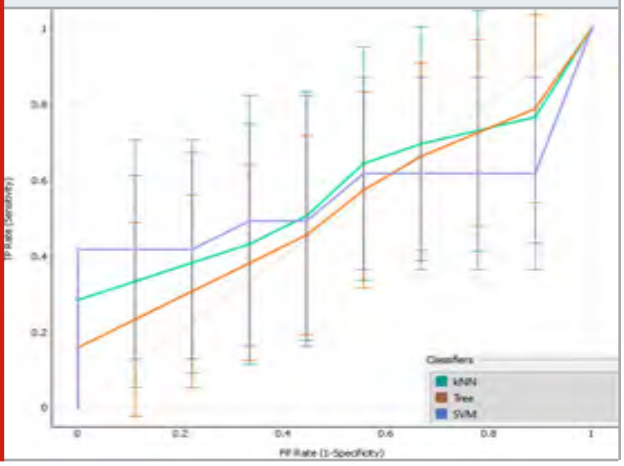
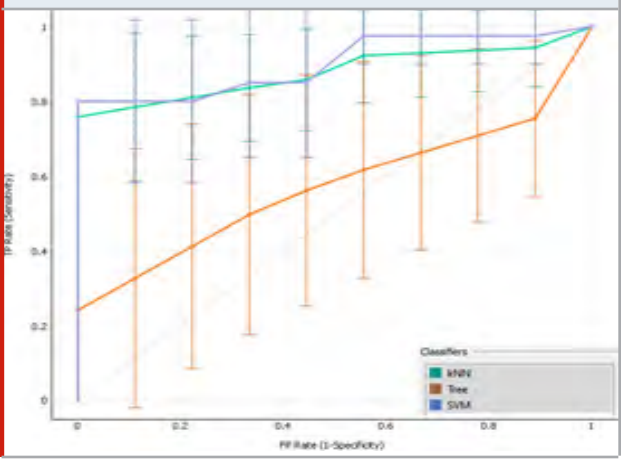


Figure 4: AUC Curve - Classification Results using CNS dataset with combination Information Gain and PCA



The graphical representation of the classification accuracy of CNS cancer dataset with various genes selected utilizing Information Gain, PCA with SVM, KNN and Decision Tree classifiers were shown in figure 4. The peak in the graph indicates highest classification accuracy of the dataset after applying the proposed method.

CONCLUSION

In the field of medical investigation, gene selection is a great concern for the researchers while predicting disease in patients. Proper selection of genes contributes to the performance improvement of the classifier model in terms of quality, accuracy, and complexity. This paper presents a various feature selection method and classification techniques applied in CNS data set. In the proposed work, ROC displays the finest subset of genes thus offering a good accuracy of classification of CNS data set. The investigational results prove that the proposed combination of gene selection method and PCA along with SVM provides improved results

when compared to other techniques. Based on the investigational findings, the suggested approach displays a good performance accuracy by using smaller number of records and selected genes.

Conflicts of Interest: The authors declare no conflicts of interest.

REFERENCES

Alexander Statnikov, Constantin F. Aliferis, Ioannis Tsamardinos, Douglas Hardinand Shawn Levy (2005), "A comprehensive evaluation of multicategoryclassification methods for microarray gene expressioncancer diagnosis", Bioinformatics, Vol. 21, No. 5, pp 631-643.

Ding C, H. Peng (2005), Minimum redundancy feature selection from microarray gene expression data, J. Bioinf. Comput. Biol. 3 (02) , 185–205.

Du D, K. Li, X. Li, M. Fei (2014), A novel forward gene selection algorithm for microarray data. Neurocomputing 133, 446–458.

Fan L, K.-L. Poh, P. Zhou (2009), A sequential feature extraction approach for Naïve Bayes classification of microarray data. Expert Syst. Appl. 36, 9919–9923.

Huang D, T.W. Chow (2005), Effective feature selection scheme using mutual information, Neuro computing 63, 325–343.

Luque-Baena R.M, D. Urda, J.L. Subirats, L. Franco, and J.M. Jerez (2013), "Analysis of Cancer Microarray Data using ConstructiveNeural Networks and Genetic Algorithms", IWBBIO, Proceedings, Canada, pp. 55-63.

NCI Brain Tumor Progress Review Group (2005) ; "Report", http.//accessible.ninds.nih.gov/find-people/groups/brain-tumor-prg/BT PRGReport.htm, February, 2005.

Rabia Aziz, C.K. Verma and Namita Srivastav (2015) , "A Weighted-SNR Feature Selection from Independent Component Subspace for NB Classification of Microarray Data", International Journal of Advanced Biotechnology and Research, Vol. 6, Issue 2, pp. 245-255.

Tan F, X. Fu, Y. Zhang (2008), A.G. Bourgeois, A genetic algorithm-based method for feature subset selection, Soft Comput. 12 (2), 111–120.

Vilda P.G. , F. Díaz, R. Martínez, R. Malutan, V. Rodellar, C.G. Puntonet (2006), Robust preprocessing of gene expression microarrays for independent component analysis. Independent Component Analysis and Blind Signal Separation, Springer, pp. 714–721.

Yang Y, S Guccione, MD Bednarski (2003), "Comparing genomic and histologic correlations to radiographic changes in tumors: A murine SCC Vll model study", Academic Radiology, vol. 10, issue. 10, pp. 1165-1175.

Zhang M L, J.M. Peña, V. Robles (2009), Feature selection for multi-label naive Bayes classification, Inf. Sci. 179 (19), 3218–3229.

Zheng C H, D.-S. Huang, L. Shang (2006), Feature selection in independent component subspace for microarray data classification. Neurocomputing 69, 2407–2410.

Biosc.Biotech.Res.Comm. Special Issue Vol 13 No 11 (2020) Pp-126-132

BBRC
Bioscience Biotechnology Research Communications

Classification and Forecasting Model for Covid -19 Disease Severities based on Medical Diagnosis using Weighted Average Dynamic Time Warping Technique

Gopalakrishnan B[1], Manikantan M[2] and Purusothaman P[3]
[1]*Associate Professor, Department of Information Technology,*
Bannari Amman Institute of Technology, Sathyamangalam, Erode, Tamilnadu, India
[2]*Associate Professor, Department of Computer Applications*
Kumaraguru College of Technology, Coimbatore, Tamilnadu, India
[3]*Assistant Professor, Department of Information Technology*
Bannari Amman Institute of Technology,Sathyamangalam Erode, Tamilnadu, India

ABSTRACT

Due to Covid-19 pandemic, mankind was affected with respect to mental and physical stress. The causes of the disease have to be analysed based on the correlation factors such as Medication, Age, Gender, physical fitness and habits. In this paper we have proposed a classification model based on weighted average dynamic time warping approach to detect the disease severity as High, Medium, and Low by considering the multi-variant dependent variables that affect the prediction of Covid -19 positive cases. We also proposed the forecasting model based on the time series exponential moving average to identify the growth of disease with respect to Age, Gender and Medical history of the covid-19 positive patient. The results are obtained by defining the correlation function to measure the disease severity in the range of High, Medium and Low. The time series analysis is done with respect to mean average disease severity and also number of positive cases. The forecasting is performed based on the age, gender and existing disorders in health. The results are analysed with other time series classification models such as weighted time wrapping to make the model fits maximum to the available input.

KEY WORDS: COVID-19 SYMPTOMS, TIME SERIES ANALYSIS, CLASSIFICATION, FORECASTING, WEIGHTED AVERAGE DYNAMIC TIME WRAPPING.

INTRODUCTION

The Coronavirus disease considers a huge disaster which not only affected the healthcare but it also affected the human life in larger proportion. In the Covid-19 pandemic, people faced more crises in their education, business, transportation, supply chain etc. People on counting the days of recovery, the cases were exponentially increased. The World Health Organisation (WHO) took a biggest initiative in assessing the people all over the world. However, there are no vaccines, the WHO has released certain guidelines and some rules for the people to be safe. As this disease caused by human-human transmission as all the people must follow social distancing. Infected people undergo certain treatment according to their severity level.

ARTICLE INFORMATION

*Corresponding Author: gopalakrishnanb@bitsathy.ac.in
Received 20th Oct 2020 Accepted after revision 11th Dec 2020
Print ISSN: 0974-6455 Online ISSN: 2321-4007 CODEN: BBRCBA

Thomson Reuters ISI Web of Science Clarivate Analytics USA and Crossref Indexed Journal

NAAS Journal Score 2020 (4.31) SJIF: 2020 (7.728)

Online Contents Available at: http//www.bbrc.in/
Doi: http://dx.doi.org/10.21786/bbrc/13.11/27

The Controlling of coronavirus disease has become a great challenge for the healthcare and also for the social workers. On further cases, they found that few are in asymptotic stages. The asymptotic condition is that the people don't have any symptoms regarding to coronavirus. As that people will easily leads to death. Technologies like Artificial Intelligence and Data Science plays an important role for analysing how severe can disease be and also in finding the disease in a shorter while. Many applications were developed in order to assess their risk level. They also recognized the Coronavirus dataset respective of different parameters.

As the cases increases exponentially, we have used the time series exponential moving average forecasting model because the analysis can be updated to any recent change in parameters and it will also provide you the quick and accurate analysis of the disease. We also used classification algorithm in order to classify the risk level into Low, Medium and High. From the dataset, we will compute a mean average on certain parameters to identify the severity of the disease and also to find the number of positive cases. In this paper, the analysis is done on age, gender and their existing disorders. By using these data, we will be able to find the risk level of a person. For accuracy, we have used the time series classification model called weighted time wrapping for best fit for the input data. The accuracy gets increase as increase in fit model percentage.

Literature Survey: Ahasan Ullah Khan et al., established a paper in which he describes how the human novel virus was initialized first in China, Wuhan in the month of December 2019.This has been the major serious situation most of the peoples are facing in the current world. There has been many serious diseases most people are getting affected. Now the Covi-19, has a vast effect on damaging the people's immune system by affecting the human population. He describes the history of the disease, prevention, and to cure the disease in the real world. Eghbal Hosseini et al., describes the Corona virus efficiency problem which is spread across the world. We proposed three solutions to solve the efficiency obstacles in the distribution process. This will minimise and reduce most of the infected people rate by lock downing the people inside the homes in order to reduce the affected rate. Many algorithms have been proposed to lessen the spread of coronavirus and they used CVA algorithms using the mathematical calculation to find the optimal solution of the graph. Quoc-Viet Pham et al., represents the Corona virus situation about Artificial Intelligence (AI) and analytics for Coronavirus.

This coronavirus has vastly connected over more than 214 nations and zones on the planet, and has fundamentally influenced each part of our everyday lives. This project aims at describing their values requesting the people about the corona virus outbreak and taking caution about the huge calamities of the coronavirus situation. In this paper the analysts with different experiences furnish the manners in which the large information about the Corona virus situation results in halting the Corona virus breakdown.

Vinay Chamola et al., describes the unparalleled eruption of the 2019 novel Covid, named as COVID-19 by means of the World health Organization (WHO), has set various administrations about the world.COVID-1 epidemic united by the anxiety of abused healthcare schemes and it has enforced a widely held countries into a state of biased or full lock down. In response to such acts, we investigation about the utilization of advances, for example, the Technologies such as IoT, AI , big data analytics and Blockchain which helps us to cure the infections of Novel coronavirus. Muhammad Adnan Shereen et al., established paper in which it explains about the coronavirus situation which is severely spreadable and virus infection caused by severe SARS-CoV-2 in China spread across the country.

In this the human transfer has been wide and there's no legally permitted medicine to obtainable to fight against the Novel coronavirus.We have analysed the serious situation of Coronavirus disease and the previous diseases SARS-CoV and MERS-CoV which deals with the pandemic situation. Sanjay Kumar et al., describes the identification and creation of sets of transferable bug datasets. Data processing, a method of discovery of noiseless appearances of massive data is one amongst such techniques that become additional widespread for treating large disease information set. The primary objective is to enhance watching systems in exaggerated conditions and UTs in Republic of India which can be terribly valued to the management, clinicians, the forces concerned in sympathetic significance of the disclose of new coronavirus (COVID-19) to enhance the government plans, pronouncements, remedial amenities (apertures, testing kits, disguises etc.) Ram Kumar Singh et al., represents the Coronavirus situation in the country which is consistently expanding from January last.

It leads to serious lock down. It plays a crucial responsibility in finding the statistics of Coronavirus disease under the prescribed conditions which have the tendency to test the features of applied math prototype which supported the classic Hot-Winters methodology as suggests that of describes the Novel coronavirus disease forecasts for Asian nation at different health conditions. An entire recovery from COVID-19 can happen solely where Associate calculable 450 days from January 2020. Hao Xu et al., proposed a system in which the block chain enabled a privacy scheme for tracing the covid-19 infected patients which senses the user information and location ID. It helps to determine the battery

performance, storage and location by minimizing the public notification of block chain for lower computing power. This will declare the user private key and declare the notification message to the blockchain user.

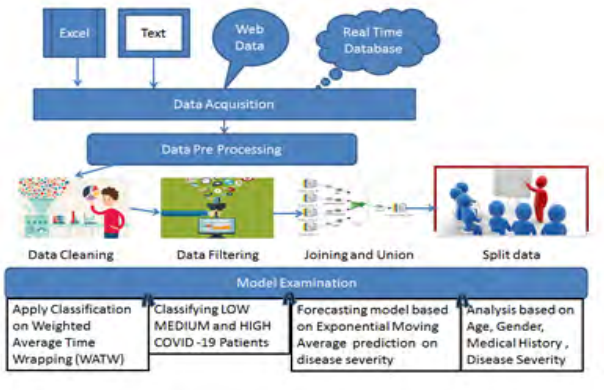
This will enhance the privacy key notifications to the user and the patient for covid-19Vasilis Z. Marmarelis et al., proposed a system in which developed a dynamic decomposition model for covid-19 which is conducted on a daily basis to show the newly confirmed cases from the early stage of covid-19 till date. They have analysed this using the time series model by forecasting methods. The graph distinguishes the infection waves using different representations. This will help the people to easily characterize the affected people by using different colour variations. The RM dynamic representation mainly characterize the mathematical parameters using Auto-Regressive Integrated Moving Average Heet savla et al. ,presented a paper in which the data's are gathered from Johns Hopkins coronavirus research center.

The data are trained using various machine learning models.This trained machine learning model will first diagnose the infection and inform it to the patient. This will be increased day by day and the number of infected cases rises accordingly. The overcoming SVM and Linear regression models helps to determine the prediction of cure cases all over the world. The KNN model will characterize the person who is infected or not. These algorithms define the monitoring and keep record of pandemic situations in our country.

PROPOSED MODEL

Overview: The proposed model consists of various steps to classify and forecast the disease severity based on the different parameters like Age, Gender, Medical history and disease severity of the persons identified through global data sources. The Proposed Process model is defined below in figure 1.

Figure 1: Proposed Model for Covid-19 classification and forecasting based time series data



The model consists of three modules

1. **Data Acquisition**
2. **Data Pre-processing**
3. **Model Examination**

Data Acquisition: The personal information on covid-19 is collected from various formats such as datasets , web content, table database and live data collection from IoT devices. The data are retrieved through governmental sources, world bank open interface, mobility data from Google and apple reports. The package has been developed to download the information from the above sources using the following code .

```
# install the package
install. Packages("COVID19");
# load the package
library("COVID19");
```

The Covid-19 data hub research committee provides unified data from world health organisations. The Package CRAN Package Covid-19 which has greater clarity and reliability in the data source that is collected with respect to the different language and geographical constraints. The WHO also provides current data related to Covid -19 for the researchers to download and upgrade the analysis based on their requirement. The data are collected from other sources like hospitals reports based on the Diagnosing SARS-COV-2 with antibodies and Diagnosing SARS-COV-2 with Nucleic-acid based tech to identify the other metrics which can be integrated to calculate the severity metric of the individual test report. The other parameters such as the patient's medical history, whether they have other medical issues like asthma , cancer, cardiology disorder , diabetes and high blood pressure etc.

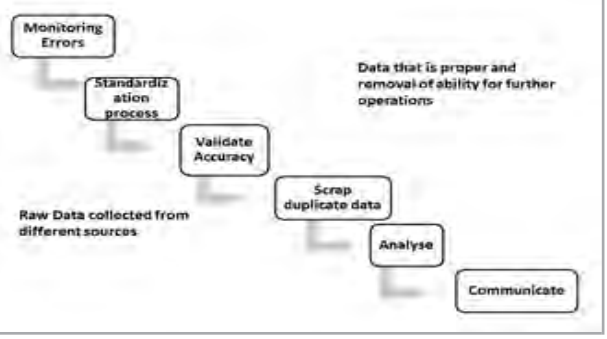
Data Pre-processing: The data pre-processing is an important aspect in the classification process using machine learning techniques. There are several steps involved in the data pre-processing that enables the data to be cleaned and perfectly matches the input requirements for the machine learning models.

Data Cleaning
Data filtering
Joining and Splitting of data
Data pivoting

Data Cleaning: The processes of deleting the unwanted information in the data source that are collected from various sources of the covid -19 disease diagnosis. The data collected contains the name and other irrelevant data that doesn't have impact on the analysis of the data set collected. In our process each table consists of the name , age, source link , journal and study in which the

data has been collected. To delete these columns from all the diagnosis tables and be able to minimize the complexity of the predictive analysis in classification. The date of diagnosis is an important factor in the time series classification models. The process of the data cleaning is done through the following steps as shown in Figure 2.

Figure 2: Flow of Data Cleaning



Data Filtering: The process of filtering is accomplished for several benefits that accomplish the data to be more consistent and error prone. The first and foremost step is to identify the dependent and independent variables in the data sources. The columns are combined to form a pivoting variable that reduces the column size. For example if the covid - 19 diagnosis data collected every day with respect to specific timing then it can be pivoted with respect to date so that every day have only one entry in the dependent variable. The measuring the error in the specific column data by considering the data value two different sources which finds the deviation in the measurement obtained from sources and the mean square error is obtained using the formula

$$MSE(\hat{y}, y) = E\left[(\hat{y} - y)^2\right] = \frac{1}{N}\sum_{n=1}^{N}(\hat{y}_n - y_n)^2$$

The next process in filtering is identifying the clone data collected from different sources and removing the duplicate information from different sources for the same instance SAR reports collected every day.

$$\text{maximum...} \frac{\left[\frac{\text{hours}}{\text{day}}\right] \times \left[\frac{\text{minutes}}{\text{hour}}\right] \times \left[\frac{\text{sechonds}}{\text{minute}}\right]}{\left[\frac{\text{sechonds}}{\text{report}}\right]} = \frac{24 \times 60 \times 60}{30} = 2880\left[\frac{\text{reports}}{\text{day}}\right]$$

The contradictory data should be identified in the data set that leads to inconsistency in the accuracy obtained by the proposed model. This can be achieved by the interpolated values between the values from different sources.

C. Model Examination: The proposed model consists of two modules

- Classification with respect to time series algorithm
- Forecasting on the severity of the disease covid-19

The classification algorithms are broadly divided into four categories

- Binary Classification
- Multi-Class Classification
- Multi-Label Classification
- Imbalanced Classification

The process of time series classification deals with the repeated pattern of the data analysis with respect to date and time field. The classification time series algorithm has been proposed as Weighted Average Dynamic Time Warping (WADTW) which can perform the average series of matrices that are generated through the repeated sequence in the linear regression prediction of the data source.

Let two division of sequence as A = { $x_1,x_2,x_3...x_n$}
B = { y_1,y_2,y_3 ... yn } where x1,x2,x3 are dependent variable and y_1,y_2,y_3 are independent variables.
Average of the x_1,x_2,x_3 are taken with respect to the weight assigned in the every iteration.

The Euclidean Distance for De (x,y) = Σ Avg (A - B)^2
The logistic weight function is defined as

$$w(a) = \frac{w_{max}}{1 + e^{-g\cdot(a - m/2)}},$$

W_{max} is the upper bound on the weight and g is the penalty level for warping.

D. Algorithm for Weighted Average Dynamic Time Wrapping

Parameters: stiffness parameter ν , penalty value λ
Let D be an n + 1 × n + 1 matrix initialised to zero.
Dist (1 , 1) ← 0
Dist (2 , 1) ← x1^2
Dist (1 , 2) ← y1^2
for a ← 2 to n + 1 do
Dist (a, 1) ← Dist (a − 1, 1) + | $x_{a-2} - x_{a-1}$ |
for b ← 2 to n + 1 do
Dist (1 , a) ← Dist (1 , b − 1) + | $y_{b-2} - y_{b-1}$ |
for a ← 2 to n + 1 do
for b ← 2 to n + 1 do
if a > 2 and b > 2 then
dist1 ← Dist (a − 1 , b − 1) + ν × | $_{a-b}$ | × 2 + | x a − 1 − y_{b-1} | + | x a − 2 − y_{b-2} |
else
dist1 ← Dist (a − 1 , b − 1) + ν × | a − b | + | x_{a-1} - y b − 1 |
if a > 2 then
dist2 ← Dist (a − 1 , b) + | $x_{a-1} - x_{a-2}$ | + λ + ν

else
dist2 ← Dist (a – 1 , b) + | x a – 1 | + λ
if b > 2 then
dist3 ← Dist (a , $_{b-1}$) + | $y_{a-1} - y_{b-2}$ | + λ + ν
else
dist3 ← Dist (a , b_{-1}) + | y_{b-1} | + λ
Dist (a , b) ← min (dist1 , dist2 , dist3)
return Dist (n + 1 , n + 1).

The data has to be splinted as training and testing for the model that predicts the severity of the disease in the required output.

E. Forecasting model to predict the severity of the particular condition:
Symbolic aggregate approximation for window prediction

SAX_Window()
Set window length ln
Set word size ws
Set alphabet size α1
Compute lookuptable
ln = length(t_s)
for all t in [0, ln – l) do
normed ts = z normalize(time series [ti : ti +l])
PAA = compute PAA (normed ts)
S = ""
for vi in PAA do
S += lookup(vi)
end for
Add S to the final representation
end for

Dividing testing and training data based on the window size
Training Time Series TrTS
Training Data TrTS ()
Set window size wi = 20
Set alphabet size α = 5
for li = min li ,wi <= Li, w+= sqrt(L) do
S = SAXwindow(raw data, li, wi,α)
Train the SEQL model from the
symbolic representation Mi = TrTS(SAX)
M[li, wi, α] = Mi
end for

Testing Time series TeTS
Testing Data TeTS ( )
Set window size wi = 20
Set alphabet size α = 5
score = 0
for l = (min li,wi <= L, wi += sqrt(L) do
sax = SAXwindow(raw data, li, wi, α)
score += M[li, wi, α].predict(SAX)
end for
Predict = sign(score)

Feature Selection
FSelection ( )
Set window size wi = 20
Set subsequence size α = 5
Set of features Fi = {}
for l = min li,wi <= L, wi += sqrt(L) do
SAX = SAX_window (raw data, li, wi, ai)
Train the SEQL model Mi = TrTS(SAX)
for all subsequence in Mi do
Fi .add(new Feature(ssi, li, wi, α))
end for
end for

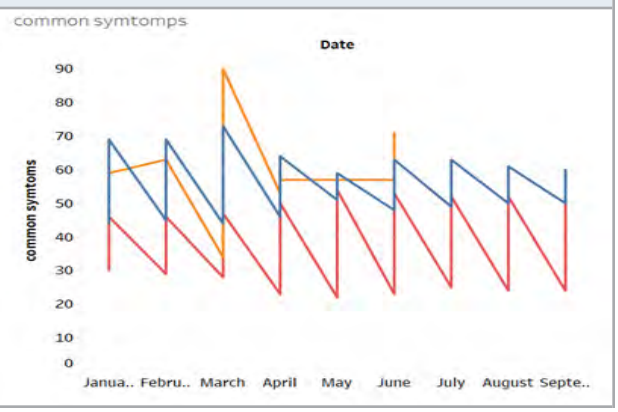
RESULT AND ANALYSIS

The process of implementing the above algorithms on the dataset collected from various sources such as open covid19 dataset. The dataset contains the more than 25 symptoms measured with date and country. The dataset has been pre-processed to eliminate the inconsistent and unwanted columns from the dataset. The analysis are made based on the following conditions

- The unsupervised clustering is performed based on Common_Symptoms (Co_S), Rarely_Symptoms (Re_S) and Critical_ Symptoms (Cr_S).
- The classification are done based on the Low, Medium and High Severity Level on the three clusters Co_S, Re_S and Cr_S,.
- The forecasting model that predicts the growth rate of the clustered symptoms Co_S, Re_S and Cr_S individually.
-

A. Cluster_symptoms time series analysis:

Figure 3: Common_Symtoms



The analysis report for the above Figure is shown in the following table: The Classication of Low , Medium and High are differentiated based on the severity levels of the Commom symptoms. In the same way Rarely_ symptomss and Critical_symptonss are classified based on Low, Medium and High with respect to time series from January – Septenber 2020 covid-19 Datasets.

Variables:	Sum of symptom:Fever
	Sum of symptom:Headache
	Sum of symptom:Pain
	Sum of symptom:Cough
Level of Detail:	Month of Date
Scaling:	Normalised

Number of Clusters:	3
Number of Points:	13872
Between-group Sum of Squares:	492.45
Within-group Sum of Squares:	437.1
Total Sum of Squares:	929.55

		Centres			
Classification	Number of Items	Sum of symptom:Fever	Sum of symptom:Headache	Sum of symptom:Pain	Sum of symptom:Cough
LOW	4369	5.4943	4.5191	36.128	7.9761
Medium	820	12.836	4.0875	30.775	12.198
High	8683	3.9904	3.7829	32.102	3.3234

B. Rarely_symptoms time series analysis.

Analysis Report: The analysis is performed based on the four different parameters for the time series classification. The P-value becomes 0. it implies the classification is perfect match with these variables.

C. Critical_symptoms time series analysis.

The analysis of critical symptoms such as nausea, fartigue, anomia, agansia as shown in figure 5.

Analysis Report

D. Forecasting of Co_S, Re_S and Cr_S based on Time Series: The forecasting was performed on the cluster model to identify which symptoms will increase in the fore coming months based on the previous information's. The following figure 6 shows the forecast values for all the symptoms and their trend analysis.

Trend Analysis: A polynomial trend model of degree 4 is computed for common symptoms given Date Month.

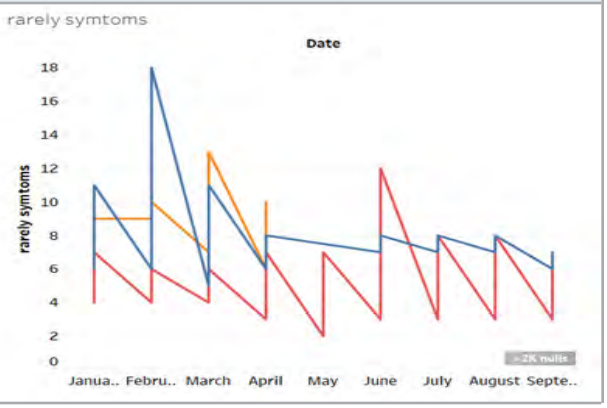
Number of modeled observations:	9
Number of filtered observations:	0
Model degrees of freedom:	5
Residual degrees of freedom (DF):	4
SSE (sum squared error):	34.4841
MSE (mean squared error):	8.62102
R-Squared:	0.793371
Standard error:	2.93616
p-value (significance):	0.110442

Since the R-Square value is 79 % which is in the acceptable line that fits the actual values

Forecasting Analysis
Options Used to Create Forecasts

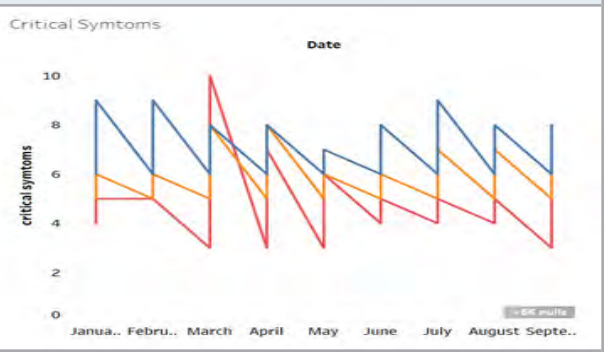
Time series:	Month of Date
Measures:	common symtoms
Forecast forward:	4 months (October 2020 – January 2021)
Forecast based on:	January 2020 – September 2020
Ignore last:	No periods ignored

Figure 4: Rarely_Symptoms vs Time Series



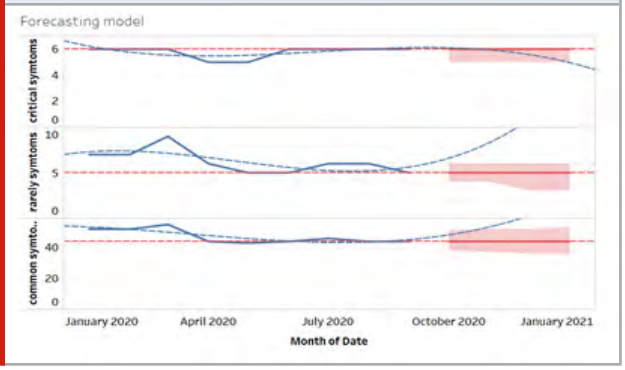
			Model		Error	
Variable	F-statistic	p-value	Sum of Squares	DF	Sum of Squares	DF
Sum ofsymptom:Sore throat	4899.0	0.0	226.5	2	285.2	12336
Sum of symptom:Shortness of breath	4306.0	0.0	115.3	2	165.2	12336
Sum of symptom:Chills	3321.0	0.0	29.45	2	54.7	12336
Sum of symptom:Diarrhea	1386.0	0.0	6.751	2	30.03	12336

Figure 5: Critical_Symptoms vs Time Series



			Model		Error	
Variable	F-statistic	p-value	Sum of Squares	DF	Sum of Squares	DF
Sum of symptom:Nausea	2991.0	0.0	95.74	2	129.2	8071
Sum of symptom:Fatigue	2935.0	0.0	100.0	2	137.5	8071
Sum of symptom:Anosmia	105.5	0.0	0.5085	2	19.46	8071
Sum of symptom:Ageusia	82.82	0.0	0.6075	2	29.6	8071

Figure 6: Forecasting vs Time Series



Since the forecasting is based on additive and multiplicative to forecast the value of MAPE which is lower 10 % then the model is acceptable.

All forecasts were computed using exponential smoothing

Common Symptoms Results

Model			Quality Metrics					Smoothing Coefficients		
Level	Trend	Season	RMSE	MAE	MASE	MAPE	AIC	Alpha	Beta	Gamma
Additive	None	None	4	3	1.12	6.2%	11	0.500	0.000	0.000

Rarely Symptoms Results

Model			Quality Metrics					Smoothing Coefficients		
Level	Trend	Season	RMSE	MAE	MASE	MAPE	AIC	Alpha	Beta	Gamma
Multiplicative	Multiplicative	None	1	1	0.68	10.6%	10	0.000	0.047	0.000

Critical Symptoms Results

Model			Quality Metrics					Smoothing Coefficients		
Level	Trend	Season	RMSE	MAE	MASE	MAPE	AIC	Alpha	Beta	Gamma
Additive	None	None	0	0	1.44	6.6%	-10	0.000	0.000	0.000

CONCLUSION

The covid -19 is an unpredictable disease due to many syndromes that affect the human mankind. It also have a property of spreading across the globe which leads to continuous deaths. This paper mainly focused on the symptoms that are more relevant to the cause of the Covid-19 results positive. The model is classified based on the severity of the certain symptoms and their impact are classified as low, medium and high with respect time series data. The proposed model also forecast the severity of the symptoms such as Common-Symptom's, Rarely_Symptom's and Crtical_Symptom's with respect to the future time line. In future it can be used to predict the consequences arise due to other diseases like Asthma, Cancer, and Diabetes. The results are analysed based on the values obtained by implementing the algorithms such as Weighted Average Dynamic Time wrapping which prominently shown better output.

REFERENCES

Ahasan Ullah Khan, Arnika Afrin Proma, Margia Akter, Md. Matiur Rahaman, Shobhan Das"A Review on Coronavirus Disease (COVID-19) Epidemic Threat for Global Health in 2020",IEEE,2014J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

Eghbal Hosseini , Kayhan Zrar Ghafoor , Ali Safaa Sadiq, Mohsen Guizani , and Ali Emrouznejad," COVID-19 Optimizer Algorithm, Modeling and Controlling of Coronavirus Distribution Process", IEEE,2020.

Hao Xu,Lei Zhang,Oluwakayode onireti,Yangfang,William J Buchanan, Muhammed Ali Imran ," BeepTrace:Blockchain -enabled Privacy preserving contact Tracing for Covid-19 Pandemic and Beyond, IEEE, 2020

Heet Savla,Vruddhi Mehta,Ramchandra Mangrulkar,"Prediction and Diagnosis of COVID-19 Using Machine Learning Algorithms", IJRTE,2020.

Muhammed Adnan shereen, Suliman khan, Abeer kazmi,Nadia Bashir,Rabeea Siddique,"Covid-19infection:Origin, transmission and Characteristics of human coronaviruses", Journal of Advanced Research, 2020.

Quoc-viet pham, Dinh C.nguyen, Thien huynh-the, Woo-joo hwang, pubudu n.Pathirana "Artificial Intelligence and Big data for coronavirus pandemic: A Survey on the State-of-the-arts", IEEE,2020.

Ram kumar singh, Martin Drews, Manuel De la sen , Manoj kumar,S.S. Singh,A.k.pandey, Prashanth kumar srivastava,Manmohan Dobriyal, Meenu Rani,Preeti kumari,Pavan kumar,"Short term statistical forecasts of Covid-19 Infections in India",IEEE,2020.

Sanjay kumar "Monitoring Novel coronaviruses Infections in India by cluster analysis",Springer,2020

Vasilis Z.Marmarelis, Fellow," Predicting Modeling of covid-19 data in the US:Adaptive phase-space approach, IEEE, 2020

Vinay Chamola,Vikas Hassija,Vatsal Gupta, Mohsen Guizani," A Comprehensive Review of the COVID-19 Pandemic and the Role of IOT, Drones, AI, Blockchain amd 5G in managing its impact",IEEE,2020.

Biosc.Biotech.Res.Comm. Special Issue Vol 13 No 11 (2020) Pp-133-137

BBRC
Bioscience Biotechnology Research Communications

Performance Comparison of Pan Tompkins and Wavelet Transform Based Ecg Feature Extraction Techniques

S N Shivappriya[1], K. Maheswari[2] and S. Sasikala[3]
[1,2,3]*Kumaraguru College of technology, Coimbatore, India*

ABSTRACT

Electrocardiogram (ECG) signals represent the heart's electrical activity in terms of P-QRS-T wave components. It is necessary to denoise and extract the components from the raw ECG signal, which is fetched from the electrodes placed on the human chest. This work compares the Pan-Tompkins and Wavelet transform Technique for extracting predominant features from the ECG signals. The fiducial points like amplitude, time period, ECG signal's onset and offset points are detected based on the windowing, thresholding Techniques. The accurate delineation takes place with appropriate scaling and wavelet functions. The various performance parameters like sensitivity, positive predictivity and accuracy are used to compare these feature extraction methods. The accurate feature extraction will give the accurate information about the heart functioning, which will improve the exact detection of the ECG signal wave components, diagnosis, and treatment.

KEY WORDS: ECG, PAN TOMPKINS, WAVELET TRANSFORM, DETECTION, DELINEATION.

INTRODUCTION

Electro Cardiogram: Electro Cardio Gram (ECG) represents the human's heart's electrical activity, recorded by skin electrode. ECG signal pattern reflects the condition of the human heart [Bonow Libby & Mann Zipes, 2006]. It is a noninvasive technique [O'Rourke, RA (Ed.) 2005], the electrodes are positioned on the human body through these electrodes the ECG signals are measured [Mathers CD et al., 2004]. If there is any variation in the morphological pattern of the ECG signal, indicates the abnormality in heart rate or rhythm, which is the reflection of the cardiac arrhythmia, these arrhythmias can be detected by analyzing from the taped ECG waveform. The physiological nature of the heart information is present on the duration and amplitude of the ECG signal. The ECG wave represents the depolarization and repolarization of the ions in the myocardial tissue [Thom Thomas Thom 2006], which forms the P-QRS-T waves.

ARTICLE INFORMATION

*****Corresponding Author: shivappriya.sn.ece@kct.ac.in
Received 20th Oct 2020 Accepted after revision 11th Dec 2020
Print ISSN: 0974-6455 Online ISSN: 2321-4007 CODEN: BBRCBA

Thomson Reuters ISI Web of Science Clarivate Analytics USA and Crossref Indexed Journal

NAAS Journal Score 2020 (4.31) SJIF: 2020 (7.728)

Online Contents Available at: http//www.bbrc.in/
Doi: http://dx.doi.org/10.21786/bbrc/13.11/28

MATERIAL AND METHODS

Databases: PhysioBank [Physionet database 2010] is a library of biomedical signals, which can be used by the research society. PhysioBank has a collection of healthy persons ECG record and holds the patients with abnormal and acute condition records like Myocardial Infraction, Bundle Branch Blocks, Ventricular Tachycardia, Ventricular and Atrail Arrhythmias etc.,

Denoising Ecg Signal: Raw ECG signal fetched from the patient body consists of the following noises:

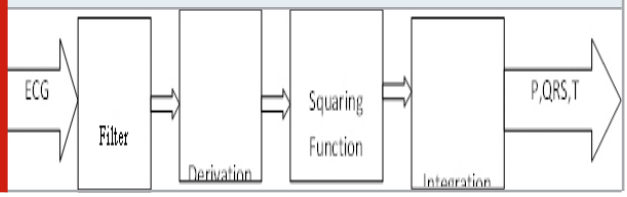
- Electrode movement noise
- Baseline wandering and drift
- Power line interference
- Muscle noise
- Channel noise
- Instrumentation noise

With Finite Impulse Response (FIR) filter, Linear filter, Infinite Impulse Response (IIR) filter, Non-linear filter and the adaptive filtering techniques are used for removing the noises present in the raw ECG signal [Raimon Jané et al., 1992]-[Seema Nayak et al., 2012]. These filters will have the drawback of time and frequency localization, which can be overcome with the Wavelet Transform based Denoising technique [Mikhled Alfaouri and Khaled Daqrouq 2008].

Feature Extraction: After Denoising the raw ECG signal from the noisy or distorted ECG signal. The extraction of the ECG signal components (P-QRS–T waves) is essential for the analysis. With the traditional rule based feature extraction technique, the features like amplitude, frequency onset and offset of the ECG wave elements are derived. [Issac Nivas et al., 2005] Shows how the efficient feature extraction technique will improve the performance of the automated detection and classification processes. This work shows that how the wavelet transform based delineation approach improves the performance of the Denoising process than the traditional Pantompkin's Algorithm Feature Extraction / Detection technique.

Pantompkin's Algorithm for Feature Extraction: Pan & Tompkins et al., (1985) is the traditional feature extraction method used for getting features from the P-QRS-T waves. After filtering process, the derivation, squaring, and integration functions used to find out peaks like Q, P and T amplitudes. In addition to these two adaptive thresholding techniques used to predict the QRS complex. In some arrythmia's there will not be any QRS complex, in those situations search back algorithm has been used to detect the P and T wave. The flow diagram of Pan Tompkins Algorithm is shown below.

Figure 1: Pan Tompkins Algorithm for ECG Segmentation Flow diagram



Band Pass (BP) filter: The concatenation of the Low and high Pass forms the Band Pass(BP) filter. The BP filter cutoff frequency is 5-15 Hz, which is the energy frequency range of QRS complex. This band pass filter not only filters the ECG components and suppresses the noises like muscle artifacts, power-line disturbance, base-line drifting etc.,

Equation (1) shows the Low pass filter Transfer Function (Trans Flp)

$$\text{Trans } F_{lp}(z)=(1-z1^{-6})\ 2\ /(1-z1^{-1})^2 \quad (1)$$

F_c =11 Hz, Gain = 36. Equation (2) shows the difference operation

$$yd(nT)= 2yd(nT- T) - yd(nT - 2T) + xd(nT) - 2xd(nT - 6T) + xd(nT -12T) \quad (2)$$

$$TFlp(z) = (1 - z1^{-32})/(1 - z1^{-1}) \quad (3)$$

$$TFhp(z) = z\ 16 - TFlp(z1)/32 \quad (4)$$

$$TFhp(z)=(-z1^{32} + 32z1^{16} -32z1^{15} + 1)/(32z1^{32} -32z1^{31}) \quad (5)$$

$$p(nT)= xd(nT - 16T) - 0.0313\ [yd(nT - T) + xd(nT) - xd(nT - 32T)] \quad (6)$$

Lower Cut off Frequency (F_c) = 5 Hz , Gain =1 and delay is 80 ms.

Derivative: ECG signal derivation shows the QRS complex slope. With the following transfer function.

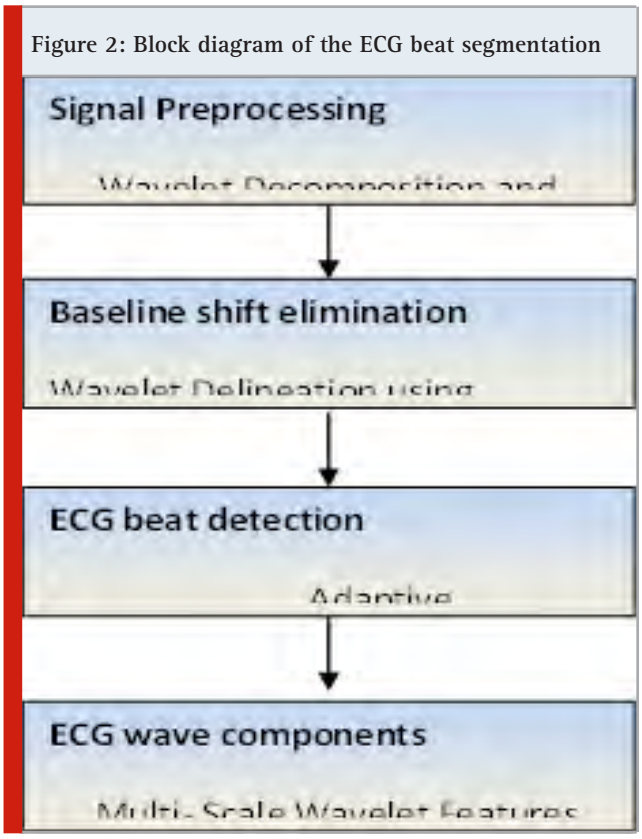
$$\text{Trans } F(z)=0.1*(2 + z11-z1^{3} -2z1^{4}) \quad (7)$$

The Trans F's difference equation is:

$$yd(nT)= (1/8) *[2xd(nT) + xd(nT - T) - xd(nT - 3T) - 2xd(nT - 4T) \quad (8)$$

Squaring: The squaring of the signal shows higher positive values.

Figure 2: Block diagram of the ECG beat segmentation



$$yd(nT)=[xd(nT)]2 \quad (9)$$

Moving integrator: The detection of R wave is not a supreme way to detect the QRS complex, in the abnormal ECG signal there may be much large amplitude and long

duration QRS complexes. In those situations, the moving integrator used to detect those wave components. The difference equation of the moving integrator is:

$$yd(nT)= (1/N)* [(xd(nT)-(N-l)T) + (xd(nT)- (N-2)T) +. . .+xd(nT)] \quad (10)$$

Width of the moving window integrator with N samples.

2.3.2 Wavelet Transform (WT) for Feature Extraction: Features play a vital role, for the evaluation of the ECG signal from the normal and abnormal patients. So, to get time, frequency and amplitude localized signal, the wavelet transform is used. Sankara Subramanian Arumugam et al., 2009 shows how the wavelet transform technique, is used for the detection of Cardiac arrhythmia signal. Figure 2 shows the ECG segmentation method the QRS complex from the P and T waves. For the MIT–BIH database the QRS complex detection rate is 99.8%. It is essential to find the timing difference between the several QRS intervals than the subjective assessment of ECG morphology [Sahambi JS et al., 1997].

ECG Delineation for Feature Extraction: S.N. Shivappriya et al., (2006) shows how the detection and delineation process takes place with the Stationary Wavelet Transform (SWT). Martinez JP (2004) shows the evaluation of various datasets like MIT-BIH Arrhythmia, European ST-T and QT database. Natalia M Arzeno et al., (2009) Obtained only (Se=99.68%, Pp=99.63%) and the largest time error. The proposed wavelet-based delineation approach, for the QRS complex detection of MIT-BIH Arrhythmia Database shows the sensitivity (Se=99.30%) and Positive Predictivity (PP=99.39%). Laguna P et al., (1985) shows the Low Pass Differentiator (LPD) approach, which will not give.

faultless T-wave end point detection, as it has Low Pass Filter and differentiator, which is simple in implementation and robust to waveform variations. The WT overcomes the drawbacks of Low Pass Differentiator (LPD), in terms of sensitivity to noise and arbitrary threshold, and provides time-frequency- amplitude localization, represents temporal features with different resolution, which is the suitable technique for analyzing the ECG signal. Wavelet Based ECG Delineator used to detect the diverse morphologies of QRS complex, P and T waves with the generalization of the detection technique. [Laguna P et al., 1985] LPD method is compared with the proposed method: the T wave and U wave detection and delineation performance is higher, particularly in the T wave end, and also locates different waves with different amplitudes in a more accurate manner.

Morphological Features: ECG morphology shows the series of deflections of heart muscle, which is away from the reference point on the ECG.

Morphological features:

1. QRS interval,
2. T wave interval,
3. P wave interval,
4. R amplitude,
5. P and T wave amplitude,
6. R and S amplitude,
7. QRS delineation interval,
8. T wave delineation interval,
9. P wave delineation interval
10. RR interval.

The proposed method uses the following medical metrics,

True Positive Rate (TPR) or Sensitivity

$$(Se) = \frac{TP}{TP+FN} X\ 100 \quad (11)$$

Positive Predictive Value (PPV) or Precision $= \frac{TP}{TP+FP} X\ 100$ (12)

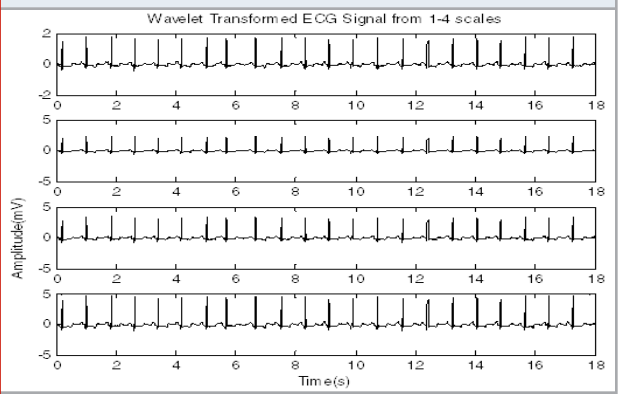
$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} X\ 100 \quad (13)$$

True Positive (TP)- truly detected measures False Negative (FN)- erroneously rejected measures False Positive (FP)- mistakenly detected measures True Negative (TN)- properly rejected measures.

RESULTS AND DISCUSSIONS

WT algorithm is utilized over the digitized and denoised ECG signal. The 21 to 25 scales are preferred which hold the predominant energy information. The QRS complex energy gets depleted when the scale is more than 24. Due to that 23 preferred for the detection of QRS complex, P and T wave is detected at 24 and 25 scale. Figure 3 shows the Wavelet Transformed and Noise removed ECG signal.

Figure 3: Shows that the Multi level (1 to 4 scale) wavelet transformed denosied ECG signal.

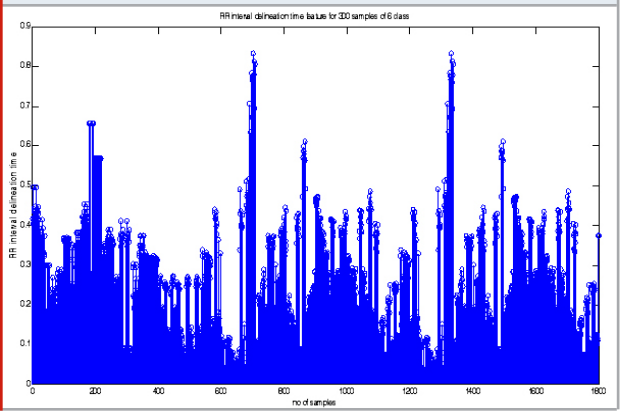


Performance Measures of Feature Extraction Process: From the MIT Data Base, QT Data Base and European Data Base 151 annotations are used for the investigation of the ECG signal. Figure 4 shows the Wavelet Transform based delineated RR intervals. Figure 5 & 6 shows the

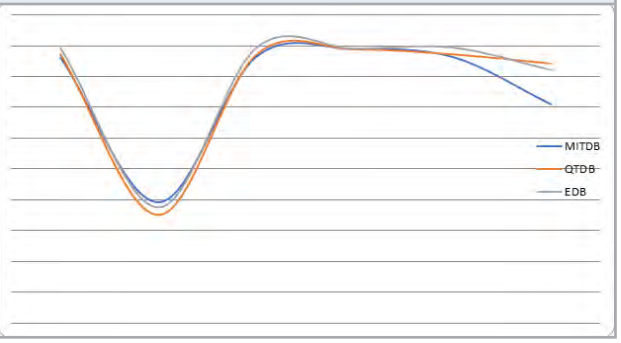
performance measure of Pan Tompkins's detection method and wavelet-based delineation method.

The Sensitivity of the proposed approach is 99.83% which is comparatively higher than the Low Pass Differentiator (LPD) method with 97.74% and Discrete Wavelet Transform (DWT) method with 99.77% sensitivity. Costas Papaloukas et al., 2002 achieved only 80.09% for the detection of T Wave episode and Se of 92.02% and PP of 93.77%. In the feature extraction process, different morphological and statistical parameters of P, QRS Complex and T wave of the ECG signals are extracted. Figure 6 shows the RR interval delineation time feature for all the individual ECG cycles of 1800 samples.

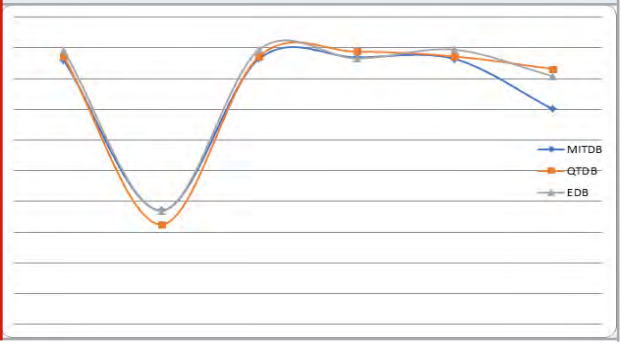
Figure 4: RR interval Delineation time Feature for 300 samples of 6 classes



Figures 5: Pan Tompkin's detection method



Figures 6: Wavelet Transform based delineation method



CONCLUSION

In ECG Monitoring System, for the analysis of several heart diseases, the detection of various features of the ECG signal is very important. This work compares the Pan Tompkins's and Wavelet Transform based feature extraction techniques for detection and delineation of ECG signal components like P-QRS-T waves. From the Feature Extraction Process, different morphological features of ECG signal are extracted : QRS detection & delineation time, R wave detection & delineation time, P wave amplitude, P wave detection & delineation time, R amplitude, S amplitude, T amplitude, T wave detection & delineation time, RR interval, Slope of ST interval. This work compares the performance of the finding and delineation process of the QRS complex, T and P waves with these measures' sensitivity and positive predictivity.

ACKNOWLEDGEMENTS

The authors thank the management and Principal of Kumaraguru College of Technology, Coimbatore for providing excellent computing facilities and encouragement. The author would also like to thank the Laboratory for Computational Physiology at MIT for providing the annotated data for this study.

REFERENCES

Bonow Libby & Mann Zipes, 2006, Heart Disease: a textbook of cardiovascular medicine, eighth edition, Saunders, Elsevier.

Costas Papaloukas, Dimitrios I Fotiadis, Aristidis Likas, Christos S Stroumbis & Lampros K. Michalis, 2002, 'Use of a novel rule-based expert system in the detection of changes in the ST segment and the T wave in long duration ECGs', Journal of Electrocardiography, vol. 35, no. 1, pp. 27-34.

Issac Nivas & Shantha Selva kumari R 2005, 'Artificial neural network based automatic cardiac abnormalities classification', Computational Intelligence and Multimedia Applications, IEEE, pp. 41-46.

Laguna P, Thakor NV, Caminal P, Jane R, Yoon H-R, Luna AB, Pan J & Tompkins WJ 1985, 'A real-time QRS detection algorithm', IEEE Transaction Biomedical Engineering, vol. 3, pp. 230-236.

Laguna P, Thakor NV, Caminal P, Jane R, Yoon H-R, Luna AB, Pan J & Tompkins WJ 1985, 'A real-time QRS detection algorithm', IEEE Transaction Biomedical Engineering, vol. 3, pp. 230-236

Mahmoodabadi SZ, Ahmadian A & Abolhasani MD 2005, 'ECG feature extraction using Daubechies wavelets', In: Proc. Fifth IASTED International Conference, pp. 343-348.

Martinez JP, Almeida R, Olmos S, Rocha AP & Laguna P 2004, 'A Wavelet-Based ECG Delineator: Evaluation on Standard Databases', IEEE Transactions on Biomedical Engineering, vol. 51, no. 4, pp.571-581.

Mathers CD, Lopez A & Stein D 2004, 'Deaths and disease

burden by cause: Global burden of disease estimates by World Bank Country Groups' Lopez AD, Mathers CD, Ezzati M, et al., editors. Washington (DC).

Mikhled Alfaouri and Khaled Daqrouq 2008. 'ECG Signal Denoising By Wavelet Transform Thresholding', American Journal of Applied Sciences 5 (3): 276-281, 2008ISSN 1546-9239

Natalia M Arzeno, Zhi-De Deng & Chi-Sang Poon 2009, 'Analysis of First-Derivative Based QRS Detection Algorithms', IEEE Trans Biomed Eng. Author manuscript; available in PMC, pp.478-484.

O'Rourke, RA (Ed.) 2005. 'Hurst's the Heart: Manual of Cardiology' (11th ed.) New York: McGraw-Hill, Medical Pub. Division

Physionet database 2010, online Available from <www. physionet.org/ physiobank /database>.

Raimon Jané, Pablo Laguna, Nitish V. Thakor & Pere Caminal 1992, 'Adaptive Baseline Wander Removal in the ECG: Comparative Analysis with Cubic Spline Technique', Computers in Cardiology, vol.2, pp.143-146.

S.N. Shivappriya, R. ShanthaSelvaKumari, T.GowriShankar. "ECG Delineation using Stationary Wavelet Transform", 2006 International Conference on Advanced

Sahambi JS, Tandon S & Bhatt RKP 1997, 'Using wavelet transforms for ECG characterization, an on-line digital signal processing system', IEEE Engineering in Medicine and Biology Magazine, vol.16, pp. 77-83.

Sankara Subramanian Arumugam, Gurusamy Gurusamy & Selvakumar Gopalasamy 2009, 'Wavelet based detection of ventricular arrhythmias with neural network classifier', JBISE, vol.2, no.6, pp.439-444.

Seema Nayak, Soni MK & Dipali Bansal 2012, 'Filtering Techniques For ECG Signal Processing', International Journal of Research in Engineering and Applied Sciences, vol.2, no. 2, pp.671-679.

Thom Thomas Thom, Nancy Haase & Wayne Rosamond, 2006, Heart disease and stroke statistics-2006 update. A report from the American heart association statistics committee and stroke statistics subcommittee, in circulation.

Biosc.Biotech.Res.Comm. Special Issue Vol 13 No 11 (2020) Pp-138-146

BBRC
Bioscience Biotechnology Research Communications

Classification of Electrocardiogram Cardiac Arrhythmia Signals Using Genetic Algorithm – Support Vector Machines

M Ramkumar[1], M Mathankumar[2] and A. Manjunathan[3]
[1]*Electronics and Communication Engineering, Sri Krishna College of Engineering and Technology, Coimbatore*
[2]*Electrical and Electronics Engineering, Kumaraguru College of Technology, Coimbatore*
[3]*Electronics and Communication Engineering, K Ramakrishnan College of Technology, Trichy*

ABSTRACT

This research study has been focused on exploring the novel approach for classifying the arrhythmia disease of cardiac muscle. The proposed methodology determines the combination of Genetic Algorithm and Support Vector Machines techniques. Initially, the feature extraction of twenty-four features has been made from the ECG waveform. The acquisition of these features has been made by partial automatic extraction from the amplitude (voltage)-time parameters of P peak, Q peak, R peak, S peak and T peak feature sets of an ECG waveform. Genetic Algorithm is utilized for improving the performance of generalization in the Support Vector Machine classifier. In order to proceed with this task, the optimization of Support Vector Machine classifier is done by performing the search of the parameter with the best value which makes tuning of its discriminate function and seeking for the best feature subset with which in response does the optimization of the fitness function associated with the classification results. Certain simulations have been carried out with the help of MATLAB software with which the results over the experiments determines the demonstration that the proposed methodology does the best classification of ECG waveforms in detecting the cardiac arrhythmias. The recorded ECG dataset has been acquired from MIT-BIH arrhythmia database and 4 different sorts of arrhythmias has been considered for performing the classification task and it is obtained with the results of 97.45% of accuracy, 95.2% of sensitivity, 95.4% of specificity and 96.23% of positive predictivity.

KEY WORDS: CARDIAC ARRHYTHMIA, ECG WAVEFORM, ECG CLASSIFICATION, FEATURE EXTRACTION, GENETIC ALGORITHM, SUPPORT VECTOR MACHINE.

INTRODUCTION

Classifying the ECG (Electrocardiogram) signals into various categories of arrhythmia disease is the tedious task of recognizing the patterns. However, ECG signal analysis is one of the most efficient methods available for making the diagnosis of cardiac heart arrhythmias. ECG arrhythmia Classification on the basis of computer approach could provide accuracy on higher rate of degree and would offer a mass potential in the screening of heart abnormalities. Achievement in doing the cardiac arrhythmia classification successfully is done by determining the ECG characteristic shape that makes the effective discrimination between the required categories of diagnosis. Certain conventional methods for the identification of typical heart beat from the ECG wave component are determined by few parameters such as area, duration and amplitude. Various statistical and morphological features are considered for inhibiting the training and testing of ECG parameters in terms of vectors to do the process of classification. Few are made on the basis of experiments done in the laboratory whereas others are made with the involvement of symptoms identified clinically.

ARTICLE INFORMATION

**Corresponding Author: mathankumarbit@gmail.com*
Received 11th Oct 2020 Accepted after revision 13th Dec 2020
Print ISSN: 0974-6455 Online ISSN: 2321-4007 CODEN: BBRCBA

Thomson Reuters ISI Web of Science Clarivate Analytics USA and Crossref Indexed Journal

NAAS Journal Score 2020 (4.31) SJIF: 2020 (7.728)

Online Contents Available at: http//www.bbrc.in/
Doi: http://dx.doi.org/10.21786/bbrc/13.11/29

In this proposed study, the ECG signals are acquired from MIT-BIH cardiac arrhythmia database. Many researchers have been widely utilizing this database for testing their different algorithms in detecting the cardiac arrhythmia and do the process of classification. The proposal of certain methods has been made for the ECG cardiac arrhythmia classification. The technique that has been presented in first study is on the basis of Fisher Linear discriminant (Acharya U R et al., 2008). The duration of R-R interval and the spatial distance in-between the existence of P peak and T peak waveforms has been perceived. On utilizing the mentioned features, the application of Fisher's Linear Discriminant has been made. The technique on the Support Vector Machine (SVM) for detecting Premature Ventricular Contraction arrhythmia is declared as more efficient algorithm when compared with ANFIS (Osowski, S, T.H. Linh, 2001). Later on, the proposal over the selection of features and the cardiac arrhythmia classification on the basis of Particle Swarm Optimization-Support Vector Machine (PSO-SVM) has been made (Chazal P., et al., 2004).

A novel approach on the basis of fuzzy-neural network has been described for classifying the ECG cardiac rhythms (Bandyopadhyay, S, S.K. Pal, 2007). In this classification mechanism, the characterization of QRS complex for the ECG waveform is determined by the polynomials of Hermite, with which the fuzzy-neural classifier is being fed by its coefficients. The proposal of Cardiac Arrhythmia detection with respect to Wavelet Transform (WT) and Independent Component Analysis (ICA) for extracting essential features has been made (Mark, R.G., et al., 1997). Later on, the proposal for classifying ECG cardiac arrhythmia by utilizing the ECG beats of huge dataset is carried out in possessing the training of neural network using the features of timing and wavelet. The identification of authors has been made in such a way that the 4th scale of dyadic WT with the wavelet of quadratic spline combined together with the post/pre interval ratio of R-R is considered as very efficient technique in differentiating and categorizing the PVC (Premature Ventricular Contraction) and Normal beats from the alternate class of ECG beats.

Even though, the development of huge techniques in classifying the ECG arrhythmias or the ECG waveforms is being carried out by various researchers, this study has also been proposed in one such way of classifying the cardiac arrhythmia by utilizing the hybrid technique of Genetic Algorithm and Support Vector Machines. The structure of this paper has been made as follows. The second section makes the description of classification methodology by Support Vector Machines (SVM). The third section determines the description of proposed technique for extracting the features in terms of feature extraction. The fourth section makes the description of the dimensionality (features) reduction. The proposed Genetic Algorithm-Support Vector Machine (GA-SVM) technique of classification has been described in fifth section. Lastly the sixth and seventh section makes the presentation over conclusion and future works to be carried on respectively.

MATERIAL AND METHODS

Classification using Support Vector Machine (SVM): In this particular section, the brief explanation and the review of 2 class SVM and the multi class Support Vector Machine classification technique has been made. Support Vector Machine (SVM) (Clifford, G.D., et al., 2006) is considered to be as very powerful and popular technique in the learning that has been created for patterns since its characteristics over the support of multi or high dimensional information and also due to the yield of best properties over generalization. Moreover, Support Vector Machines have many applications over recognizing the patterns, applications over data mining such as categorizing the text (Acharya U R et al., 2017, Homaeinezhad, M.R., et al., 2011), recognition of phoneme (Banerjee, S., Mitra, M., 2014), detection of three-dimensional object (Liu, T., et al., 2016), classification of images (Mitra, M., Samanta, R.K., 2013) and the field of bio-informatics (Martis, R.J., et al., 2013b). At the initial stage, the formulation of Support Vector Machine has been made for the problems of classifying binary (2-class). The extension over this particular technique has been made to the problems of multi-level class which might neither be straightforward nor be the isolated one. DAG-SVM is one of the techniques with which the proposal has been made for the extension of Support Vector Machine Classifier for supporting the classification of multi-class domain (Thomas, M., et al., 2015).

Formulation of Binary Support Vector Machine

Let $X = \{(x_i, y_i)\}_{i=1}^{n}$ be considered as the set of training samples n, with which $x_i \in \mathcal{R}^m$ is the sample of m-dimension in the space of input and $y_i \epsilon \{-1,1\}$ is denoted as the label class for the sample x_i. Support Vector Machine determines the finding over the optimal separating hyperplane (OSH) with the minimal errors resulted from the classification. The hyperplane separation in the linear form is being represented as in equation (1).

$$f(x) = w^T x + b \tag{1}$$

Where w denotes the vector of weight and b denotes the bias respectively. The hyperplane with optimal characteristics could be acquired by resolving the problem of optimization, where ξ_i is determined as the variable of slack for acquiring the soft margin whereas the C variable determines the control for the variables of slack. The value for the margin of separation enhances by providing the decrement for the C value.

In the SVM, a hyperplane which is optimum in nature has been acquired by enhancing the capability of generalization of Support Vector Machine by itself. However, if the data associated with the training are not separable linearly, the acquired classifier might not possess high ability of generalization although optimal determination of the hyperplanes are being carried out. For enhancing the separability in the linear fashion, the

mapping for the original space of input has been made with the space of the dot product with high dimension represented as the feature space. Now utilizing the function of vector which is non-linear $\varphi(x) = (\varphi_1(x), \dots \dots \varphi_1(x))^T$ with which the mapping is made over the input vector x with m-dimension into the feature space of l-dimension, the OSH (Optimum Separation Hyperplane) in the feature vector space is denoted by equation (2).

$$f(x) = w^T \varphi(x) + b \tag{2}$$

The decision function that has been designated for the test information is given by equation (3).

$$D(x) = sign(w^T \varphi(x) + b \tag{3}$$

$$\begin{aligned} &Minimize\ \tfrac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \\ &subject\ to\ y_i(w^T\varphi(x_i) + b) \geq 1 - \xi_i \\ &\xi_i \geq 0, i = 1, \dots \dots \dots, n \end{aligned} \tag{4}$$

Figure 1: Functional Block Diagram of proposed classification of ECG arrhythmia

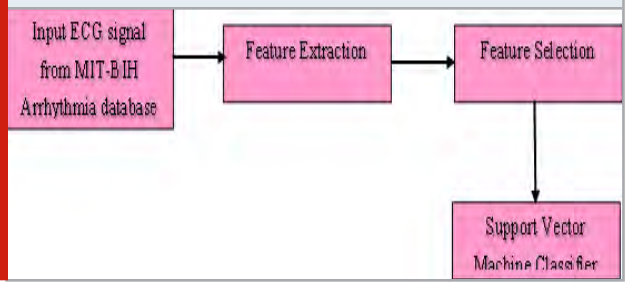
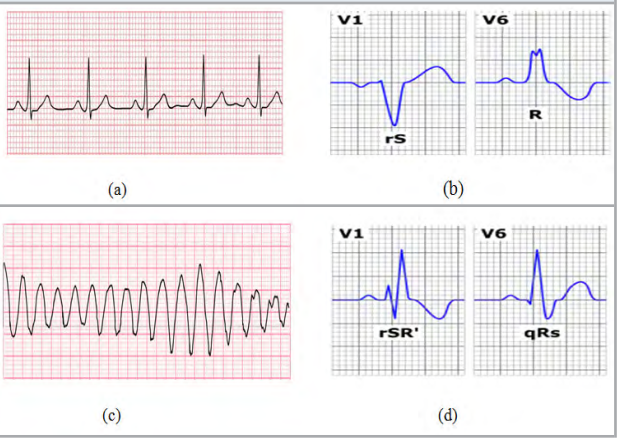


Figure 2: Representation of sample signal a) Normal Sinus Rhythm b) Left Bundle Branch Block Beat c) Premature Ventricular Contraction d) Right Bundle Branch Block Beat



Support Vector Machine with Multi-Class: As the description made earlier Support Vector Machines are the classifiers with binary intrinsic characteristics. However, the ECG waveform classification makes the involvement for more than dual classes. For facing this particular problem, an adaptation for the total number of strategies involved under the classification of multiclass has been made [15] Daubechies, I., 1998). Among those strategies, the most essential strategies are one-against-all (OAA) strategies and one-against-one (OAO) strategy. The construction for the One Against One (OAO) has been made with (n(n-1))/2 decision functions for most of the classification pair combinations. The results over the experimentation has indicated that the One-Against-One is best sustainable and suitable for the practical usage. In this proposed study, the process of OAO has been utilized for the multi class classification of ECG signal (Takeuchi, K., Collier, N., 2003).

Methodology for the Extraction and Selection of features: In this particular section, the discussion has been made in such a way that the characteristics of the ECG features that has been extracted and the design procedure for extracting the ECG features has been explained. Figure 1 depicts the functional block diagram of the proposed classification mechanism of ECG arrhythmia. The flowchart for the algorithm execution has been discussed in the following chapter.

Description of Dataset: The conduction of the experiments has been made over the ECG waveform as the raw base signal which is being taken for signal classification. In most of the research studies the realization over the ECG signal classification has been acquired from MIT-BIH cardiac arrhythmia database (Wang, T.-Y., Chiang, H.-M., 2007) from the physionet and it is available online with open access. It has been widely utilized for evaluating the performance of various classifiers. The database is totally consisting of 48 recordings with which each individual recording is being made with 30 minutes of time duration. The recording of each data is made is made in dual channels and it is denoted in terms of second modified limb lead and sixth modified limb lead.

Particularly, the type of ECG beat which has been considered in our proposed study is being determined by the following classes. They are Normal Sinus Rhythm (NSR), Left Bundle Branch Block Beat (LBBB), Premature Ventricular Contraction (PVC) and Right Bundle Branch Block Beat (LBBB). The sample representations of those ECG beat signals has been depicted in figure 2. The selection of those beats is made from 101, 104, 106, 109, 118, 124, 207, 214, 219, 221, 231 and 234 as shown table 1.

Reduction of Noise: In the initial stage of extracting the features, the performance over the wavelet transform has been made in order to enhance the reduction of noises. The WT (Wavelet Transform) permits the non-stationary signal processing such as Electrocardiogram signal. The representation of raw ECG signal and the Filtered ECG signal has been depicted in figure 3 and figure 4 respectively.

Description of Features: For each of the ECG signal, totally 19 features which are temporal in its characteristics such as interval of R-R, interval of P-Q, interval of P-R, interval of P-T and added to that the recognition of 3 morphological features has been made. Manually, the extraction of these features has been made for each ECG beat and stored into an isolated vector. The tag of each vector has been made with one of its 4 labels like NSR, PVC, LBBB and RBBB respectively.

Table 1. Description of Dataset and Usage of numbers in simulation

Class Number	Example of record acquired from MIT-BIH	Total count of ECG beats used	Type of Beat
1	101, 104	245	Normal Sinus Rhythm (NSR)
2	106, 109	108	Premature Ventricular Contraction (PVC)
3	207, 214	602	Left Bundle Branch Block Beat (LBBB)
4	118, 124	448	Right Bundle Branch Block Beat (RBBB)

Figure 3: Raw acquired ECG signal

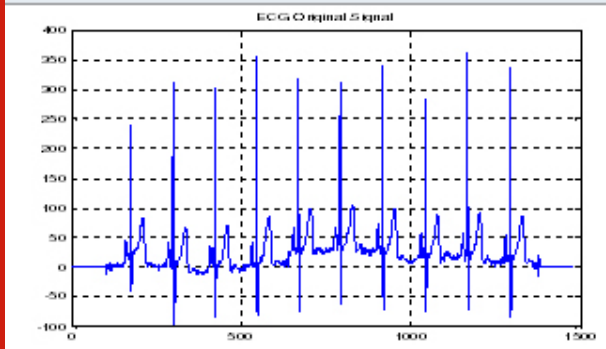
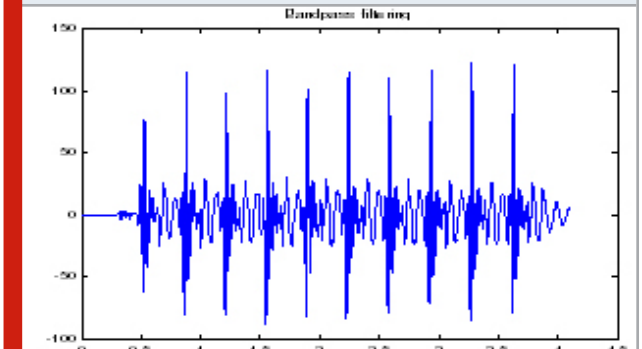
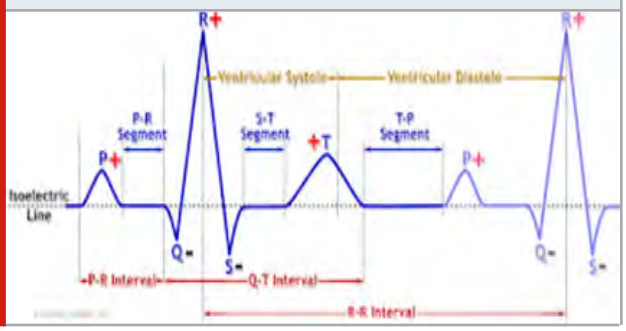


Figure 4: Filtered ECG signal



The extraction of the features has been made inclusive of the P peak, Q peak, R peak, S peak and T peak along with the time interval of R-R, S-T and P-T and interval of time over each of five various features from the preceding feature such as R-S, S-T and Q-R as denoted in figure 5 and also it is inclusive of the feature's voltage difference that has been denoted in terms of E(Q)-E(S). Consideration of one more feature has been made in terms of voltage and time of R-R. The features which has been described is being represented in the following table 2. Y(R) denotes the R position of ECG wave component and W(R) denotes the position's value in that ECG component.

Figure 5: Representation of Features for ECG Wave Component



The computation of the minimum and the maximum values of ECG beats in the signal has been yielded with the formation of 3 features which are morphological in its characteristics. Scaling over each beat in the ECG signal has been made by utilizing the subsequent formula with which its range of each individual signal relies between 0 and 1.

$$h(t) = \frac{h(t) - \min(t)}{\max(t) - \min(t)} \quad (5)$$

The maximum voltage and the minimum voltage in between the initial and the subsequent feature of R is being computerized and followed with the performance of normalization process [0 1]. As determined earlier, the consideration over the percentage has been meant with the value which is larger than 0.2, 0.5 and 0.8 as 3 various features. Out of 22 features, 6 features are being represented as the basic features and they are as follows: R1, R2, P,Q,T and S and the remaining are represented as the features of derivatives. The calculations of the features that has been derived are made utilizing the basic features through the procedure of semi automation. The

suggestion for the initial and the subsequent point of R to the expert is being made by utilizing the minimum-maximum based algorithm. Then the differentiation made by the experts for the appropriate points has been made in terms of P, Q, R, S and T.

Reduction of Features: Many researches in the area of selection of features and data analysis determines the suggestion that not every feature is being utilized for classification process (Bandyopadhyay, S, S.K. Pal, 2007, Mark, R.G., et al., 1997, Osowski, S, T.H. Linh, 2001). As on the case of contrary, few features might function as noisy parameters and hence it results with the reduction in the accuracy of classification. In this proposed study, two various approaches on the reduction of features has been accomplished. This proposed mechanism has shown that the technique on the basis of meta-heuristic has yielded best performance for the arrhythmia classification on ECG signal when compared to the method of statistical analysis (Bandyopadhyay, S, S.K. Pal, 2007).

Table 2. Description of features that has been utilized in the simulation

Feature Number	Description of Features	Feature Number	Description of Features
1	Y(R1)	11	Y(R2)
2	W(R1)	12	W(R2)
3	Y(S)	13	Y(R2)-Y(R1)
4	W(S)	14	W(R2)-W(R1)
5	Y(T)	15	Y(S)-Y(R1)
6	W(T)	16	Y(T)-Y(S)
7	Y(P)	17	Y(P)-Y(T)
8	W(P)	18	Y(Q)-Y(P)
9	Y(Q)	19	Y(R2)-Y(Q)
10	W(Q)		

Principal Component Analysis: Principal Component Analysis is determined as the probabilistic technique for making the reduction in the data dimension (Wang, T.-Y., Chiang, H.-M., 2007). It establishes the selection on the variable set which are not correlated with one another and parallelly each variable is declared as the linear combination of originated variables. The derivatives of the principal components are made from the original information such that the initial principal component calculates for the maximum proportionality of the original informational set variance inclusive of orthogonal components which are subsequently present that has also been accounted for the peak proportionality for the balance variance. The successive steps for the execution of Principal Component Analysis has been mentioned as follows.

a. Determine the computation for the information mean vector
b. Determine the computation for the information covariance matrix
c. Determine the computation for the eigen matrix and the eigen value of the matrix of covariance
d. Formulation of the components utilizing the covariance matrix eigen vectors as the coefficients of weights.

It is sustained to be advertised that the classifier of the Principal Component Analysis (PCA) establishes its performance well for the entire informational datasets rather the evaluation of the performance could not be attained well for few groups of informational sets (Bandyopadhyay, S, S.K. Pal, 2007). Very few research studies (Ahmad, A.M., et al., 2013, Chang, C.C., Lin, C.J. 2011, Kressel, U.H.-G, 1999, Wang, T.-Y., Chiang, H.-M., 2007) has shown that the Principal Component Analysis (PCA) is not the standard technique for making the analysis over the non-linear information. It seems to be that the noisy information existence, abnormal range of features for which some are ranging between [0 1] whereas the other features are ranging between [0 10000], minimum variance of few essential features are the major reasons for the minimal performance of Principal Component Analysis. Table 5 determines the result over experimentation by Principal Component Analysis.

Genetic Algorithm: Genetic Algorithm is one of the optimized techniques that is being utilized for the reduction of features. Few alternate techniques of optimization with meta-heuristic characteristics like simulated annealing tabu search and the strategies of evolution are also declared as the candidates for enhancing this purpose. The demonstration over Genetic Algorithm has been made for converging into the partially optimized solution for many difficult and diverse problems as the stochastic and the powerful tools on the basis of natural evolution principle (Ebrahimzadeh, A. Khazaee, 2010). In most of the applications it is being utilized for the dimensionality reduction of features and the weighting of features (Vaseghi, S.V., 2008). The steps for the algorithm in executing GA has been mentioned as follows in Table 3.

The initial step in any of the genetic algorithm is to make the definition of the encoding process for allowing the description of a potential solution as a part of numerical vector and for making an attempt to make the randomized generation of population. The brief descriptions on the operators of GA has been made as follows.

a. Selection: The process of selection directly picks the individuals from the existing population on the basis of

fitness parameters from each chromosome (Hsu, C.W. and C.-J. Lin, 2002).

b. Recombination: The main responsibility of the operation of crossover is to enable the creation of current individuals from the previous ones. Crossover is most probably determined as the statistical process which makes the exchange of data between few individuals of parent for making the generation of new individuals as child.

c. Mutation: The application of mutation is being done to single individual and makes the production of modified child mutant.

d. Function of Fitness: The main responsibility of the fitness function is to make the measurement of the solution's quality.

Genetic Algorithm-Support Vector Machine: In this particular section, the description over the Genetic Algorithm-Support Vector Machine for classifying the cardiac arrhythmias is being made. In this system, the main aim is to make the selection process for the feature subset automatically in order to optimize the classifier of Support Vector Machine. The procedure has been described as follows in figure 6.

Set up of Genetic Algorithm: The initial step in the genetic algorithm is to make the definition for the procedure of encoding which permits the description of any valid or strengthened solution as the vector of numerical value. The length of the vector has been chosen with the value of 23 with each individual component either saturates with the value of 1 or 0 which has been selected for the process of encoding. The synonym for the component that possess zero value is that the omission of the subsequent feature is being made and the synonym for the component that possess one value is vice versa. In the following experimentations, the original population is possessed with fifty chromosomes which has been selected randomly.

Table 3. Pseudo Code for Genetic Algorithm

	Algorithm Steps	Genetic Algorithm Description
Input: Training Information Output: Required Features	0th Step	Initialization of parameters (e.g. Size of population, rate of crossover, rate of mutation and the maximum count of population generation).
	1st step	Random Creation of initial population (P(0)).
	2nd step	Evaluation of current population (Computation of fitness for entire chromosomes)
	3rd step	While (satisfaction of termination condition is not being made, step 4 to 8 has to be proceeded.
	4th step	Performance of selection over P(t) from P(t+1) has to be made.
	5th step	Performance of mutation and crossover has to be made from the recombination of P(t)
	6th step	Evaluation of current population has to be made in order to establish the computation of all fitness chromosomes.
	7th step	Proceed with t=t+1
	8th step	To be proceeded with step 3

The process of swapping and the selection of roulette wheel has been utilized for the mutations and the crossover operations respectively. The operation of swap changes the position of dual samples randomly. The probability mutation parameter has to be selected with the value of 0.1. The selection for the fitness function choice is very essential since it is on the basis of evaluation made by genetic algorithm, the best of solution has been given to each candidate for establishing the design of Support Vector Machine system of classification. In this proposed study the exploration for the correction rate in performing the classification of ECG signal is being made.

Classification of SVM with GA: The description of the procedure for the classification system on the basis of Support Vector Machine is as follows.

Step 1: Generation of initial random population of determined size of 50

Step 2: For the population of each individual chromosomes, perform the training of

$\frac{n(n-1)}{2}$ Support Vector Machine Classifiers.

Step 3: Utilizing the multiclass SVM technique (OAO) the computation for the fitness of each individual

chromosomes is being made (feature subset).

Step 4: Selection of few individuals directly has to be made from the existing population on the basis of fitness values and make the regeneration of upcoming individuals from the previous ones.

Step 5: If the maximum count of iteration is not attained, then it has to be proceeded from 2nd step.

Step 6: Selection of chromosomes with the best values of fitness has been made as the desired feature subset.

Step 7: Classification process of ECG signals with the trained Support Vector Machines.

Figure 6: Approach of Genetic Algorithm-Support Vector Machine

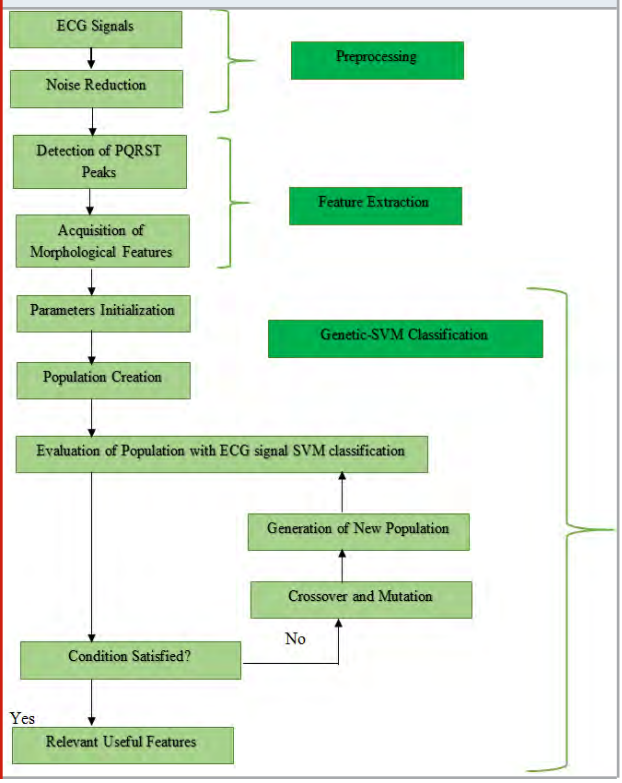


Table 4. Classification of ECG Arrhythmia using SVM with Linear and Polynomial Kernel

S No	P, LR	P, LL	P, N	LR, LL	LR, N	LL, N	Overall
1	97.96	99.99	51.25	99.45	55.56	68.74	78.83
2	92.63	98.21	54.26	99.45	53.24	68.74	77.76

RESULTS AND DISCUSSION

For evaluating the proposed technique, more than seventy five percent of informational data which has been acquired from MIT-BIH arrhythmia database are being utilized to train the system of composition and the remaining are utilized for making the evaluation. In the initial state of experiment, the stage of selection is being omitted and the application of Support Vector Machine Classifier is being directly made to the entire space of original features. The results over the experimentation has been represented in table 4, 5 and 6.

Table 5. Classification of ECG Arrhythmia using PCA_SVM with Linear and Polynomial Kernel

S No	P, LR	P, LL	P, N	LR, LL	LR, N	LL, N	Overall
1	98.24	98.65	54.72	99.54	55.68	66.82	78.94
2	99.32	99.45	55.45	97.28	55.68	66.82	79

Table 6. Classification of ECG Arrhythmia using GA_SVM with Linear and Polynomial Kernel

S No	P, LR	P, LL	P, N	LR, LL	LR, N	LL, N	Overall
1	96.26	98.45	99.26	99.99	89.75	98.11	96.97
2	96.26	90.22	86.15	99.99	74.62	95.41	90.44

Figure 7: Performance evaluation of linear kernels with three different classification algorithms

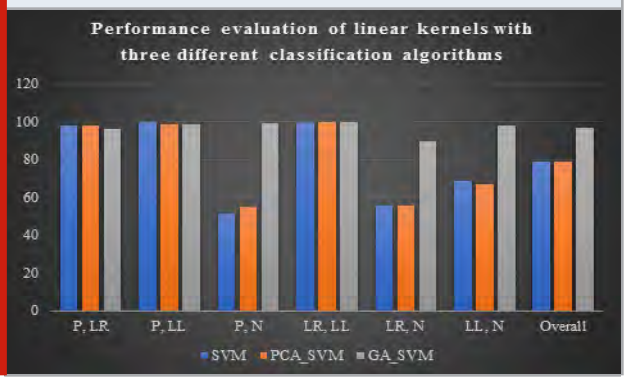


Figure 8: Performance evaluation of Polynomial kernels with three different classification algorithms

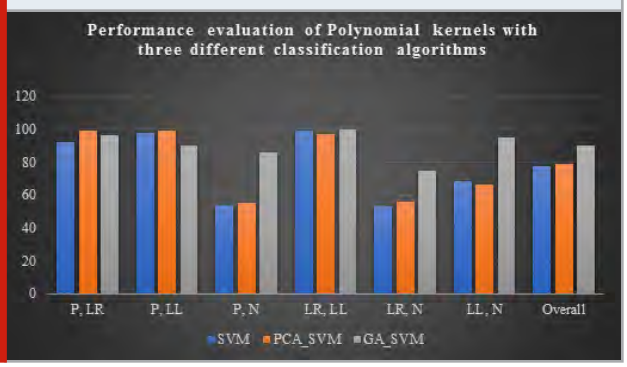
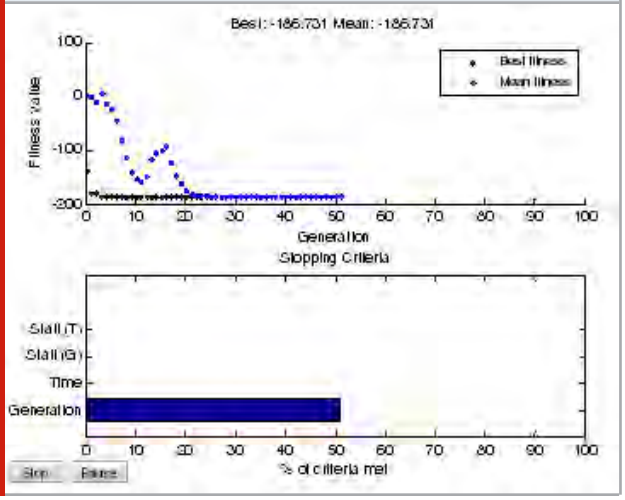


Figure 9: Evaluation between the fitness value and the population generation



In the next part of experimentation, the proposed Genetic Algorithm-Support Vector Machine Classifier system is being utilized to the best feature subset which could optimize the classifier of SVM. The conclusion is made in such a way that the GA-SVM possessed with linear kernel generally performs well and it has been provided with 96.97% of classification accuracy in diagnosing the cardiac arrhythmias. Figure 7 shows the performance evaluation of linear kernels with three different classification algorithms and Figure 8 shows the performance evaluation of Polynomial kernels with three different classification algorithms. The simulation window which determines the evaluation between the fitness value and the generation has been depicted in figure 9.

CONCLUSION AND FUTURE WORK

In this proposed study, the proposal of a novel technique on the basis of Genetic Algorithm-Support Vector Machine for the selection of features and the classification of ECG signals has been made. The results over the experimentation has proved that the methodology for the right selection of features enhances the classification accuracy and quality. This happens and exists since few features might act as noise component, and degrades the classification accuracy. A simple and efficient GA-SVM technique that has employed for ECG classification is being presented with the datasets that has been acquired from MIT-BIH arrhythmia database. The future work includes the enhancement in deep learning techniques for classification by the determination of feature selection strategies.

REFERENCES

Acharya U R, et al., (2008). Automatic identification of cardiac health using modeling techniques: a comparative study. Inform. Sci. 178 Pages 4571–4582.

Acharya, U.R., Fujita, H., Lih, O.S., Hagiwara, Y., Tan, J.H., Adam, M., (2017). Automated detection of arrhythmias using different intervals of tachycardia ECG segments with convolutional neural network. Inform. Sci. 405 Pages 81–90.

Ahmad, A.M., Khan, G.M., Mahmud, S.A., (2013). Classification of Arrhythmia Types Using Cartesian Genetic Programming Evolved Artificial Neural Networks. In: Iliadis L., Papadopoulos H., Jayne C. (Eds.,) Engineering Applications of Neural Networks. EANN 2013. Communications in Computer and Information Science, vol. 383, Springer, Berlin, Heidelberg.

Bandyopadhyay, S, S.K. Pal, (2007). Classification and Learning Using Genetic Algorithms, Springer-Verlag Berlin Heidelberg.

Banerjee, S., Mitra, M., (2014). Application of cross wavelet transform for ECG pattern analysis and classification. IEEE Trans. Instrum. Meas. 63 (2) Pages 326–333.

Chazal, P., M. O'Dwyer, R.B. Reilly, (2004). Automatic classification of heartbeats using ECG morphology and heartbeat interval features, IEEE Trans. Biomed. Eng. 51 Pages 1196–1206.

Chang, C.C., Lin, C.J. (2011). LIBSVM: a library for support vector machines. ACM

Transactions on Intelligent Systems and Technology 2 Pages 27:1–27:27.

Clifford, G.D., F. Azuaje, P.E. McShary, (2006). Advanced Methods and Tools for ECG Data Analysis, Artech House, Norwood, MA.

Daubechies, I., (1998). Orthonormal bases of compactly wavelets. Commun Pure Appl Math 41 Pages 909–996.

Ebrahimzadeh, A. Khazaee, (2010). Detection of premature ventricular contractions using MLP neural networks: a comparative study, Measurement 43 Pages 103–112.

Homaeinezhad, M.R., Tavakkoli1, E., Ghaffari1, A, (2011). Discrete Wavelet-based Fuzzy Network Architecture for ECG Rhythm-Type Recognition: Feature Extraction and Clustering Oriented Tuning of Fuzzy Inference System. In: International Journal of Signal Processing, Image Processing and Pattern Recognition, vol.4, no. 3, September.

Hsu, C.W., C.-J. Lin, (2002). A comparison of methods for multiclass support vector machines," IEEE Trans. Neural Netw. 13(2) Pages 415–425.

Kressel, U.H.-G, (1999). Pairwise classification and support vector machines. In: Advances

in Kernel Methods, Pages 255–268. MIT Press, Cambridge.

Liu, T., Si, Y., Wen, D., Zang, M., Lang, L., (2016). Dictionary learning for VQ feature extraction in ECG beats classification. Expert Syst. Appl. 53 Pages 129–137.

Mark, R.G., G.B. Moody, MIT–BIH Arrhythmia Database 1997 [Online]. Available: http://ecg.mit.edu/dbinfo.html.

Martis, R.J., Acharya, U.R., Lim, C.M., Suric, J.S., (2013b). Characterization of ECG beats from cardiac arrhythmia using discrete cosine transform in PCA framework. Knowl.-Based Syst. 45 Pages 76–82.

Mitra, M., Samanta, R.K., (2013). Cardiac arrhythmia classification using neural networks with selected features. Proced. Technol. 10 Pages 76–84.

Osowski, S, T.H. Linh, (2001) ECG beat recognition using fuzzy hybrid neural network, IEEE Trans. Biomed. Eng. 48 Pages 1265–1271.

Takeuchi, K., Collier, N., (2003). Bio-Medical Entity Extraction using Support Vector Machines. Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine, Pages 57-64.

Thomas, M., Das, M.K., Ari, S., (2015). Automatic ECG arrhythmia classification using dual tree complex wavelet based features. Int. J. Electron. Commun. 69 (4) Pages 715–721.

Vaseghi, S.V., (2008). Advanced Digital Signal Processing and Noise Reduction, 4th edition, John Wiley & Sons.

Wang, T.-Y., Chiang, H.-M., (2007). Fuzzy support vector machine for multi-class text categorization. Information Process and Management, 43 Pages 914–929.

Biosc.Biotech.Res.Comm. Special Issue Vol 13 No 11 (2020) Pp-147-152

BBRC
Bioscience Biotechnology Research Communications

Predicting Metamorphic Changes In Parkinson's Disease Patients Using Machine Learning Algorithms

G.Prema Arokia Mary[1], N.Suganthi[2], M.S.Hema[3], M.Hari Dharshini[4], K.Vaishaali[5], M.Monika Sri[6]

[1]*Assistant Professor, Department of Information Technology, Kumaraguru College of Technology, Coimbatore, India.*

[2]*Professor, Department of Computer Science and Engineering, Kumaraguru College of Technology, Coimbatore, India.*

[3]*Associate Professor, Department of Information Technology, Anurag University, Hyderabad, India.*

[4,5,6]*Student, Department of Information Technology, Kumaraguru College of Technology, Coimbatore, India.*

ABSTRACT

Parkinson's disease is a nervous disorder mainly it affects the motor activities of the human body. Manifestations start step by step; at later point it becomes the greatest obstacle to do our day to today activities. Individuals influenced with Parkinson's ailment should go through lifestyle changes and enthusiastic changes like dozing issues, disposition swings, stultification, and skin issues. The proposed methodology is to analyse the proportion of metamorphic changes of a person affected by Parkinson's disease using machine learning techniques. Principal Component Analysis (PCA), recurrent neural network and logistic regression algorithms are used for prediction. The accuracy, precision, recall and F1 measure is used to assess the performance of the prediction algorithms. The dataset includes activities of daily living which from PPMI (Parkinson's Progression Markers Initiative) was taken for experimentation. Logistic regression can predict metamorphic changes with a higher accuracy of 92% for sleep dataset and 95% for Olfactory(smell) dataset when compared to other two algorithms.

KEY WORDS: HALLUCINATIONS, LOGISTIC REGRESSION, METAMORPHIC CHANGES, NORMALIZATION, PCA, PREDICTION, RNN, TRANSFORMATION.

INTRODUCTION

Parkinson's Disease (PD) is a progressive nervous system disorder which affects the movement of the human beings. PD occurs due to lack of dopamine chemical in the human brain (Sontheimer H, 2015). Dopamine is a contraction of 3,4 dihydroxy phenethylamine. It comprises about 80% of the catecholamine content. It is also synthesized in plants and animals. It functions as a neurotransmitter in the brain (G.Prema Arokia Mary et al.., 2020). It is a chemical which sends signals to nerve cells and is released by neurons. Association network of protein builds the hereditary network collaboration of PD. There is no separate test for PD prediction. The PD is predicted based on the symptom's tests and clinical trials. Symptoms starts gradually in earlier stages. Early stages may be mild and unnoticed. Parkinson's disease signs and symptoms are different for everyone. Tremors

ARTICLE INFORMATION

*Corresponding Author: premaarokiamary.g.it@kct.ac.in
Received 5th Oct 2020 Accepted after revision 26th Nov 2020
Print ISSN: 0974-6455 Online ISSN: 2321-4007 CODEN: BBRCBA

Thomson Reuters ISI Web of Science Clarivate Analytics USA and Crossref Indexed Journal

NAAS Journal Score 2020 (4.31) SJIF: 2020 (7.728)

Online Contents Available at: http//www.bbrc.in/
Doi: http://dx.doi.org/10.21786/bbrc/13.11/30

are the most common one, but it also causes stiffness, slowing of movement and voice changes.

People usually develop this disease at the age of 60 or above. But now days young adults also affected with PD due to various factors such as environmental triggers, exposure to toxins like pesticides, smoking, head injury etc.,. Genetics also play a major role in causing PD. Young Onset Parkinson's disease (YOPD) occurs in people younger than 50 years. Youngsters also have symptoms similar to aged people like tremors in hands, arms, legs, jaw and face, rigid muscles, bradykinesia, impaired postures and balance, loss of automatic movements, speech changes etc., (Suwijn, Sven R. et al.., 2020). People with YOPD may experience non-motor symptoms including depression, sleep disturbances, anxiety, difficulty in swallowing, distorted sense of smell, unintentional writing, weight loss, stultification etc., (Feigl, Beatrix et al.., 2020, Tremblay, Cécilia et al.., 2020). The PD stages are divided into five different stages.

Symptoms of earlier stage or stage one is tremor, experiencing changes in walking and facial expressions. In stage two, people with PD experience rigidity and other movement symptoms on both sides of the body. In third stage, people face with postural disability which means movements becomes very slow, but one can manage by their own without any assistance. In stage four, people can severely affect by movement disorder. The person cannot be able to do their daily tasks, so they require a nursing care for all the activities. The fifth stage is said to be an advanced stage person become bedridden. People may experience hallucinations and delusions (Huang, Xuemei et al.., 2020).

Melancholy happens in roughly 40% of patients with Parkinson's illness. Sleep disturbances are the foremost characteristics of the non-motor issues of Parkinson's disease (PD) and the prevalence with advancing disease become greater. There are various causes of sleep disturbance in PD, and it has several factors that contributes to many patients. These can be widely classified into those which includes nocturnal sleep and daytime illustrations. The primary manifestation of PD is excessive daytime sleepiness which reflects the constructed areas which affected by the neurodegenerative process. The mental concomitant in Parkinson's disease is anxiety disturbances which confers to significant impairments in cognitive, functional, motor areas (Cummings, J. L., 1992).

It results in decreased quality of life, greater levels of care dependency and rises caregiver burden. Many of the people with PD would experience generalized anxiety disorder, anxiety attacks and social anxiety disorders. Most people who have PD, experienced a difficulty in swallowing which is called dysphagia, it can happen at any stage of Parkinson Disease (DeMaagd G, Philip A, 2015). Symptoms which include in swallowing difficulties are coughing or throat clearing during eating and drinking and difficulty while swallowing certain foods or liquids. Aspiration pneumonia is the cause of death in PD. People who experience the ability to smell which was called as dysosmia. It is an early sign of Parkinson (Laperle, A.H., Sances, S., Yucer, N, 2020). Modelling tools are used to study the disease and give prognostication of symptoms. To understand and identify the symptoms of the disease, quantitative studies and models suggested by some authors. The amalgamation of Parkinson's different data from several sources which are multiplex in nature and provide opportunities to study the premature stages of the patients, track the development and offering solutions. Heterogeneous data are handled using Big data analytics (S. Kanagaraj et al.., 2019).

The main contribution of the proposed methodology is to predict the metamorphic changes of the PD patients. The performance of the prediction algorithms is compared. This paper has been proceeded in such a way : apart from introduction along with literature section, second section consists of the proposed methodology, dataset information, data pre-processing and the implementation of three algorithms like PCA, RNN and Logistic Regression. Results are discussed in section three. The conclusion is presented in section four.

MATERIAL AND METHODS

Proposed Methodology: The proposed methodology for PD prediction using machine learning algorithms. Three algorithms are used for prediction namely Recurrent Neural Networks, Logistic Regression and Principal Component Analysis. The steps involved in the proposed methodology are importing libraries, loading the Data, data pre-processing, splitting the data as train data and test data, prediction algorithm implementation and performance assessment such as accuracy, precision, recall and f1_score calculation. Two datasets i.e. sleep disorder and olfactory (smell) datasets is taken for implementation.

Dataset Information: The dataset is taken from PPMI (Parkinson's Progression Markers Initiative). PPMI is a milestone observational clinical investigation to exhaustively assess partners of critical enthusiasm utilizing progressed imaging, biologic inspecting and clinical and social evaluations to recognize biomarkers of Parkinson's disease movement. The informational collection for dozing problem and olfactory smell is taken. The informational index contains 8367 information for Parkinson malady influenced individuals.

Data Pre-processing: Data pre-processing is an important step in machine learning that helps to strengthen the aspects of data to extract the meaningful insights from the data. It is a technique of converting the raw data to quality data. Normally raw data is incomplete, inconsistent, or it has some error or outliers. The steps involved in data pre-processing is: select the dataset, import all the crucial libraries, Import the dataset, data cleaning, data reduction and splitting. In data cleaning, the missing values, null values are identified and

removed. The data inconsistencies also removed in data cleaning. The data set has a greater number of features. All the features are not required for prediction. The irrelevant features may degrade the performance of the prediction algorithm. The required feature selection is done in data reduction. Next comes data splitting, here the data is divided into two sets namely training set and test set. The training set is used to build the model. The test set is used to check performance of the model (K. Saranya et al.., 2014).

Algorithm Introduction

Principal Component Analysis: Principal component analysis (PCA) is a technique in which data with high dimensional are lowered to low dimension by selecting the most important features that capture maximum information about the dataset (S. Sehgal et al.., 2014). Often People think that PCA takes only the selected features and discards the remaining features. But it actually builds the new set of properties based on the older ones. PCA act as powerful tool in analysing the data. The main advantage of this algorithm reduces with a smaller number of features (Rao, C. Radhakrishna, 2020). First normalization process is done so that a dataset is normalized before applying PCA algorithm. Covariance Matrix (CM) is calculated as:

$$CM= [Va[Z1]\ \ Cv[Z1,Z2]\ \ Cv[Z2,Z1]\ \ Va[Z2]\ ] \quad (1)$$

$$Va[Z1]=Cv[Z1,Z2]\ \text{ and } Va[Z2] \quad (2)$$

Eigenvalues and eigenvectors are as calculated to generate CM, λ is the eigenvalue of matrix H

$$D(\lambda I1-H)=0 \quad (3)$$

where,
I1 denotes dimensions of identity matrix are equal
D denotes determinant of a matrix

For every eigenvalue λ , a relating eigenvector v1,

$$(\lambda I1-H)v1=0 \quad (4)$$

The components are chosen, and Feature vector (FV) is formed as,

$$FV=(eig1,eig2) \quad (5)$$

where,

eig1, eig2 are eigenvalues

Principal Components can be formed as follows:

$$NewData=F^{VT}\times ScaledData^{T} \quad (6)$$

where,

New Data is a matrix which consists principal components

Scaled Data is an original dataset of scaled version

T is transpose of matrix

After implementing PCA algorithm, the performance measures like accuracy, recall, precision, F1 score are to be calculated.

Recurrent Neural Network: A Recurrent Neural Network (RNN) is a type of artificial neural networks where the interconnections from one node to another node constructs a directed graph. It is derived from feedforward neural network policy. RNN algorithm gives predictive results in sequential data. RNN algorithm used in developing the models which does similar activity of neurons in the human brain. Some applications of RNN are Robot control, Speech recognition, Grammar learning etc., also used by Google's voice search. It perfectly suits for machine learning problems and its the first algorithm which recollects the input. For the past few years, it is one of the best algorithms which has seen excellent results in deep learning. It consists of cyclic connections which make the algorithm a more efficient and powerful tool to model the sequential data. It achieved a great success while demonstrating prediction tasks like handwriting recognition and language modelling. It is performed and obtained great results in sequence labelling (Z. Tang, et al.., 2016).

First of all, independent activations are converted into dependent activations by giving similar loads to each layer. By this conversion intricacy of parameters can be decreased. The last output is always remembered, because previous outputs are given as input to the next hidden layer. Likewise, all layers are joined together, and weights remains same for every layer and it is converted into a single recurrent layer (Salehinejad, Hojjat et al.., 2017). By feeding the initial data to the network, current state can be calculated with the help of multiple outputs which taken from the previous state. It can have many steps according to the problem and once everything is done, the output is calculated which is done in final state. Then comparison between output and actual output (target output) takes place and weight gets updated. Now RNN is trained. At time t, it produces output and the parameter affected was available at time t + 1. It has a remarkable representation so that it keeps the information about the past steps. In such a way, it keeps the recent past and present inputs to produce the output for the data. The current state is calculated by the formula:

$$K_{\alpha}=f(K(_{a-1}),L_a) \quad (7)$$

whereas,
K_a denotes current state
K_{a-1} denotes previous state
L_a denotes input state n

Activation function is applied by the formula (tanh):

$$K_{\alpha}=\tanh(Y_{kk}\,K_{\{a-1\}}+Y_{lk}\,L_a) \quad (8)$$

whereas,
Y_{kk} denotes weight at recurrent neuron
Y_{lk} denotes weight at input neuron

The result Zt is calculated by using the formula:

$$Z_t = Y_{kz} K_a \quad (9)$$

whereas,
Ykz denotes weight at output layer

Logistic Regression Algorithm: Logistic regression is one of the famous algorithms to forecast dichotomous values. It predicts the probability of the outcomes (Peng, Joanne et al.., 2002). The nature of target variable is dichotomous, which means that there would be only two possible classes. In general, the target variable is in binary form and the desired outcome is either 0 or 1. Logistic Regression is one of the simplest ML algorithms that can be used for various classification problems (X. Zou et al.., 2019). The general equation to calculate linear regression is:

$$s(Ex(M))=A+Bi1+Ci2 \quad (10)$$

where,
s() is link function
Ex(M) is expectation of target variable.

A + B*i1 + C*i2 is linear predictor.

From this logistic regression equation can be derived and the equation is:

$$log\ log\left(\frac{p}{1-p}\right) = B_0 + B(Age) \quad (11)$$

Table 1. Performance comparison chart of Accuracy, Recall, Precision, F1_Score values for PCA, Logistic regression & RNN algorithm

Performance Evaluation	PCA Algorithm		Logistic Regression Algorithm		RNN Algorithm	
	Sleep Dataset	Smell Dataset	Sleep Dataset	Smell Dataset	Sleep Dataset	Smell Dataset
Accuracy	83%	80%	92%	95%	89%	86%
Recall	78%	76%	90%	95%	87%	85%
Precision	80%	76%	87%	90%	85%	86%
F1 Score	79%	74%	85%	91%	84%	82%

After training phase, test dataset prediction takes place. Once the prediction phase successfully completed the performance measures like accuracy, confusion matrix, precision, F1 score are calculated.

RESULTS AND DISCUSSION

Sleep and smell dataset are taken from PPMI repository. The attributes chosen to predict sleep disorder are Vivid dreams, Aggressive Dreams, Movement awaken and sleep disturbances and for Olfactory(smell) the attributes of scent levels from 1 to 5 and the total percentage are considered. Total 8367 instances taken for implementation. The language used for implementation is Python. In data pre-processing all the null values and missing records are removed. All the attributes are taken for prediction. Principal Component Analysis, Logistic Regression and in Recurrent Neural Network prediction algorithms are implemented. The accuracy, precision, F1 score and recall for sleep and smell dataset in each algorithm is given in table 1.

Accuracy(A) is determined by using the formula,

$$A = \frac{Q+P}{Q+P+R+S} \quad (12)$$

Precision(P) is determined by using the formula,

$$P = \frac{P}{P+R} \quad (13)$$

Recall (R) is determined by using the formula,

$$R = \frac{P}{P+S} \quad (14)$$

F1 score(F) is computed as,

$$F = 2 * \left(\frac{P*R}{P+R}\right) \quad (15)$$

Whereas,
P is said to be true positive
Q is said to be true negative
R is said to be false positive
S is said to be false negative

The logistic regression gives better performance in both sleep dataset and smell dataset when compared to other two algorithms which is shown in Figure 2.

Figure 1: Graph depicts the Sleep and Smell data by Performance Evaluation of PCA,Sleep data(skyblue colour),Smell data(darkblue colour) where, X-axis shows the accuracy,Precision,F1-Score,Recall and Y-axis shows the percentage values.

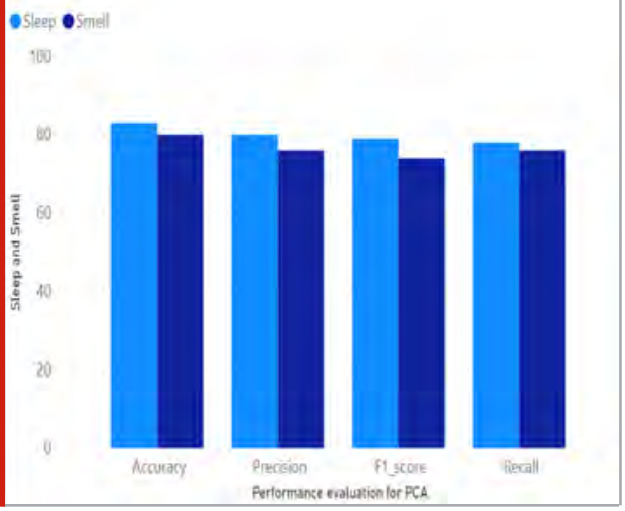


Figure 2: Graph depicts the Sleep and Smell data by Performance Evaluation of Logistic Regression Sleep data(orange colour),Smell data(yellow colour) where, X-axis shows the accuracy,Precision,F1-Score,Recall and Y-axis shows the percentage values.

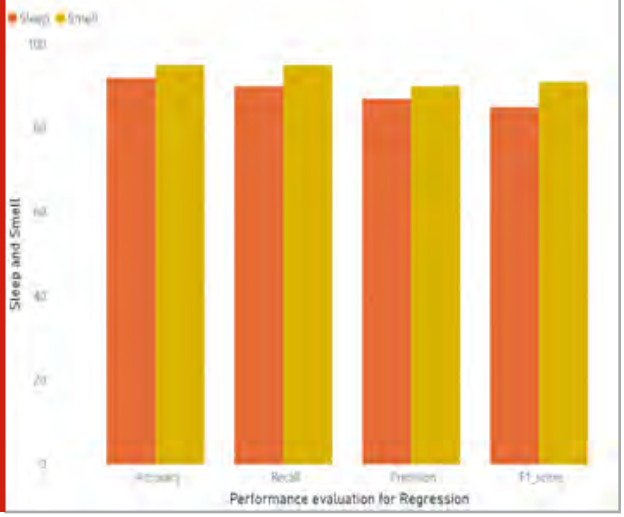
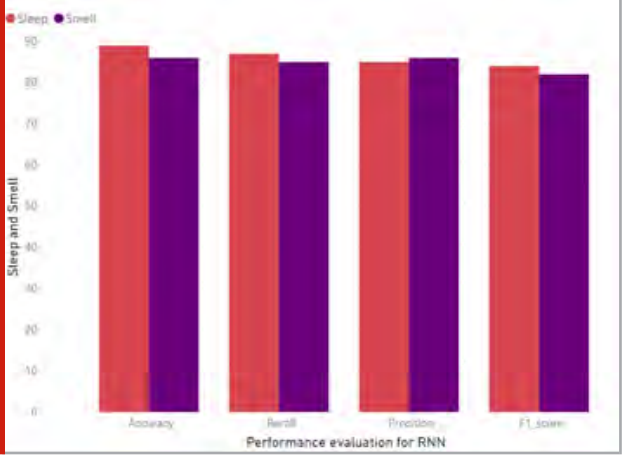


Figure 3: Graph depicts the Sleep and Smell data by Performance Evaluation of RNN Sleep data(maroon colour),Smell data(purple colour) where, X-axis shows the accuracy,Precision,F1-Score,Recall and Y-axis shows the percentage values.



CONCLUSION

The prediction of PD using machine leaning techniques is proposed and implemented. Three algorithms such as principal component analysis, recurrent neural network and logistic regression algorithms are implemented. The performance of these three algorithms are compared. The logistic regression algorithm has highest accuracy of 92% for sleep dataset and 95% for smell dataset.

REFERENCES

Cummings J. L. (1992). 'Depression and Parkinson's disease: a review'. The American journal of psychiatry, 149(4), 443–454. https://doi.org/10.1176/ajp.149.4.443

DeMaagd, G., & Philip, A. (2015). 'Parkinson's Disease and Its Management: Part 1: Disease Entity, Risk Factors, Pathophysiology, Clinical Presentation, and Diagnosis'. P & T : a peer-reviewed journal for formulary management, 40(8), 504–532.

Feigl B, Dumpala S, Kerr GK, Zele AJ (2020) 'Melanopsin Cell Dysfunction is Involved in Sleep Disruption in Parkinson's Disease'. J Parkinsons Dis ,10(4) 1467-1476. doi:10.3233/jpd-202178. PMID: 32986681.

Huang, Xuemei et al., (2020), 'The D 1/D 5 Dopamine Partial Agonist PF-06412562 in Advanced-Stage Parkinson's Disease: A Feasibility Study', J Parkinsons Dis, 10(4) 1515-1527. doi:10.3233/jpd-202188. PMID: 32986682.1 – 13

Kanagaraj. S,. Hema. M.S, Nageswara Gupta. M. (2019) 'Machine Learning Techniques for Prediction of Parkinson's Disease using Big Data', Volume-8 Issue-10, International Journal of Innovative Technology and Exploring Engineering (IJITEE).

Laperle, A.H., Sances, S., Yucer, N. et al., (2020) 'iPSC modelling of young-onset Parkinson's disease reveals a molecular signature of disease and novel therapeutic candidates',Nature Medicine,26(2), 289–299, https://doi.org/10.1038/s41591-019-0739-1.

Peng, Joanne & Lee, Kuk & Ingersoll, Gary. (2002) 'An Introduction to Logistic Regression Analysis and Reporting', Journal of Educational Research – J EDUC RES. 96. 3-14. 10.1080/00220670209598786.

Prema Arokia Mary. G, Saru Priya. K, Suganthi. N,

Sathyavathi. S. (2020), 'Dopamine Analysis Using Multiple Machine Learning Techniques', International Journal of Advanced Science and Technology, 29(05), 10220 – 10227.

Rao, Radhakrishna .C. (2020) 'The Use and Interpretation of Principal Component Analysis in Applied Research', The Indian Journal of Statistics, JSTOR, Sankhyā. vol. 26, no. 4, 1964, pp. 329–358.

Salehinejad, H., Baarbe, J., Sankar, S., Barfett, J., Colak, E., & Valaee, S. (2018). 'Recent Advances in Recurrent Neural Networks'. ArXiv, abs/1801.01078.

Saranya, K. & Hema, M.S. & Chandramathi, S.. (2015). 'Data fusion in ontology based data integration. 2014 International Conference on Information Communication and Embedded Systems', ICICES 2014. 10.1109/ICICES.2014.7033792.

Sehgal, S., Singh, H., Agarwal, M., Bhasker, V., & Shantanu (2014). Data analysis using principal component analysis. 2014 International Conference on Medical Imaging, m-Health and Emerging Communication Systems (MedCom), 45-48.

Sontheimer. H (2015) Parkinson Disease. Diseases of the Nervous System: Academic Press 133-64.

Suwijn, S. R., Samim, H., Eggers, C., Espay, A. J., Fox, S., Lang, A. E., Samuel, M., Silverdale, M., Verschuur, C., Dijk, J. M., Verberne, H. J., Booij, J., & de Bie, R. (2020) 'Value of Clinical Signs in Identifying Patients with Scans without Evidence of Dopaminergic Deficit (SWEDD)', Journal of Parkinson's disease, 10(4), 1561–1569. https://doi.org/10.3233/JPD-202090.

Tang. Z, Wang. D & Zhang. Z. (2016)'Recurrent neural network training with dark knowledge transfer', IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai.

Tremblay, C., Iravani, B., Aubry Lafontaine, É., Steffener, J., Fischmeister, F., Lundström, J. N., & Frasnelli, J. (2020). Parkinson's Disease Affects Functional Connectivity within the Olfactory-Trigeminal Network. Journal of Parkinson's disease, 10(4), 1587–1600. https://doi.org/10.3233/JPD-202062.

Zou. X, Hu. Y, Tian .Z & Shen. K. (2019) 'Logistic Regression Model Optimization and Case Analysis',IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT), Dalian, China, pp. 135-139, doi:10.1109/ICCSNT47585.2019.8962457.