Asian Journal of Social Science and Management Technology
ISSN: 2313-7410
Volume 4, Issue 2, March-April, 2022
Available at www.ajssmt.com

Machine Learning Applications Related to Medical Predictions

Chetanpal Singh

Holmes Institute, Melbourne Australia,

Abstract: It is crucial to identify diseases at the earliest possible stage to render prompt and relevant treatments immediately. It can be a matter of life and death because a late diagnosis can make a disease fatal and incurable. Many instances have been reported where humans have erred in detecting certain abnormalities prematurely because of their inability to comprehend certain medical data. It necessitates using different Machine Learning (ML) processes to analyze complex data, medical images, and reports to address the issue mentioned above. Decision-making based on these computational data is inherently more accurate and has a widespread application in medical science in general. Moreover, these machine-supported approaches help identify hidden abnormalities or patterns that human may otherwise overlook, which makes the collected data even more crucial. Machine Learning is a major tool that takes its cues from the Internet of Things (IoT) and delivers useful insights based on a handful of datasets. The algorithms used for ML are highly varied, and the predictive results identified from it may change with varying datasets themselves. Thus, it can have an overall impact on the entire decision-making process. Since the healthcare sector is sensitive, it is imperative to understand these algorithms and how they are implemented to interpret them. The paper's main aim is to highlight the different ML algorithms that have a role in interpreting medical data and predicting diseases. An in-depth analysis also reveals the shortcomings of certain algorithms. Therefore, it is critical to identify the type of algorithm to be employed for a certain dataset.

Index Terms: Machine Learning, Disease Diagnostics, Classification Algorithm, Supervised Machine Learning, Disease Prediction, Healthcare, Classification

1. INTRODUCTION

One of the most important utilizations of a health prediction model is to identify the outpatients who need immediate medical attention and who can be shifted to a lesser important section. They play an important role in streamlining the patient flow appropriately so that no major issues arise related to it. Certain emergencies are the main scenarios that increase the demands of healthcare services, and it is determined by the outpatient demand and the number of ambulances available [1]. It can lead to a situation where certain hospitals become overburdened with outpatients while some facilities remain relatively empty. The role of IoT in this situation is to create the virtual connection between the computers of the area which
can even facilitate physical communication. Microprocessor chips of the latest innovations are used to gather and disperse important information.

Artificial Intelligence (AI) depends on several concepts related to Machine Learning, Language Processing, Computer Vision and Deep Learning [2]. It consists of a system that is computerized and mimics humans in terms of thinking, perceiving, and operation [3]. The decision-making is dependent on the interpretation of past data that undergoes probabilistic, statistical, and optimization techniques. The applications of this process in various disciplines like intrusion recognition in networks, detection of the behavior of customers during purchase, and detection of fraud related to credit cards have been well-documented. The supervised learning

approach has been used in most of these applications and they have been known to be successful as well. The prediction models are used to interpret the unlabeled examples in this method [4]. Therefore, the hypothesis that is generated here is that the model can be utilized by doctors as a tool for the accurate and efficient diagnosis of disease.

A. Research Objective

This literature aims to look at the various trends that are prevalent in the ML model that can be used for premature detection of disease using important metrics for performances [5]. Some of the common algorithms that have been employed for the same are Decision Trees (DT), K-Nearest Neighbor (KNN) and Naive Bayes (NB). Certain diseases related to the heart, brain, kidney, and breasts will be analyzed using these algorithms. Many methodologies will be applied for the evaluation of these diseases and the model that performs best concerning a certain disease will be ultimately chosen so that the prediction becomes accurate.

B. Research Motivation

An important aspect for the appropriate treatment of a disease is to identify it as early as possible and Machine Learning can play a critical role in it [6]. It analyses previous data that is already available to accurately predict the diagnosis of a disease. Several prevalent methods are efficient enough to facilitate diagnosis. A model can be developed using machine learning algorithms to bring about a diagnosis at an early stage which can be imperative to reduce the death rate of any pathology. It is due to this reason that doctors these days choose to rely on machine learning for the correct disease prediction [7].

The examples that are already available in training are used as a learning dataset and the patterns are derived out of it. Inferences are also developed using the different prediction models. It also helps to set up certain models for classification that can point out a disease comprehensively [8]. The main criteria are to identify the patterns that generally remain hidden, and it can play an instrumental role in helping the healthcare sector take the next step towards evolution. Patients suffering from diseases related to vital organs like the brain, kidney, and liver can benefit from it largely. Several classifiers can play an important role in this aspect. Some of them include the Random Forest, K-nearest, and Naive Bayes [9].

Certain important hotspots help to understand the disease prediction model more accurately using supervised Machine Learning algorithms and the result of the study can help the researchers in devising a more comprehensive outcome.

C. Research Gap

Even though using Machine learning for disease prediction has gained a lot of popularity in recent times, there is still a lot of room for progress and research so that it can be universally accepted. In fact, there is very little literature that considers these learning algorithms and reviews them which hampers its credibility quite comprehensively.

2. LITERATURE REVIEW

Machine Learning is a recently evolving branch of Artificial Intelligence that enables a machine to make a judgment after thinking like a human without any supervision whatsoever [10]. The machines are not programmed, and the process of thinking takes place automatically. Computers form an integral part of Machine Learning, and they have access to huge datasets which are used for the learning itself [11].

There is input and output data and the prediction of it is known as supervised learning. It is quite dissimilar to unsupervised learning which only contains the input data for reference and the structure is more one-sided [12]. Another type of learning is called the semi-supervised one and it is a mixture of both the above-mentioned technique. The labeled, as well as unlabeled data, are used for this purpose [13]. Reinforcement learning involves a method of learning through an interaction with the environment so that the desired actions are rewarded, and the undesired ones are punished [14]. The main area of interest lies in making accurate diagnoses that may look different. Deep Learning is another concept that contains layers of perception that need to interact with each other for proper results [15].

Previously, it was the expertise of the doctor along with their intuition and knowledge, along with available statistics, that was used to predict diseases. However, it also led to errors in practices, biases, and errors which impeded the quality of treatment [4]. The situation has changed recently as the data available is more widespread and a robust approach has been taken using Machine Learning protocol for the same function. On conducting a thorough literature review, it was revealed that each study utilized several ML algorithms. This made the comparison and identification of the most appropriate one difficult. This study primarily focuses on comparing the performance of these different algorithms to facilitate the process of choosing the ideal one for a number of diseases.

Machine learning application in medical predictions

The purpose of Machine Learning has been to detect disease at a nascent stage so that the treatment is provided at the earliest possible which improves the chances of curing the patient completely and reducing the mortality rate. Several diseases have been identified but a varying range of algorithms have been used for the same. The purpose of this section is to list down a few diseases which can be treated better when detected early, the algorithm of Machine Learning that is used to do it, and the predictive features associated with it. The comparison of these features will be done in detail in the discussion portion of the article along with suggestions to improve them.

A. Breast Diseases

Breast Cancer is responsible for the maximum number of deaths from cancer for women. An important method for reducing its morbidity is by detecting it early through mammograms, Magnetic Imaging Resonance (MRI), biopsy, and ultrasound. The tumors that cause breast cancer are of two types, benign and malignant. Differentiating between both clinically is difficult for a physician and the malignant counterpart is known to be more fatal [16]. The role of Machine Learning becomes extremely crucial here as it collects and analyses the data that is available to automatically detect breast cancer without being explicitly guided to do so.

Various Machine Learning Techniques have gained a lot of ground in recent times in the premature detection of breast cancer. The analysis is mainly done in 3 stages which include preprocessing stage, classification, and extraction of features [17]. The most important stage among these 3 is the feature extraction stage as it helps to distinguish between the benign and malignant tumors effectively. The segmentation process is used after the extraction based on certain properties like size, depth, coarseness, and smoothness of the tumor.

The best way to take out reliable information from images is by converting it into binary. However, there has been a change in opinion in recent times as the conversion into binary sometimes results in the loss of some crucial features. Thus, having the image in greyscale itself has been advocated. Discrete Wavelet Transformation (DWT) method is used for changing time-domain images to frequency ones [18]. This is done by using four different matrices known as the vertical, horizontal, diagonal, and coefficient matrix. The values that are extracted from it are used by the ML algorithm.

[19] used different ML algorithms like Bayesian Networks, RF, and SVM to detect breast cancer prematurely. The dataset was obtained from Wisconsin Repository and various features like the accuracy, precision, ROC graph, and recall were used for the analysis. An important classifier called the k-fold method of validation was used which implied that the value that is chosen for K is 10. As far as accuracy, recall, and precision were concerned, SVM was found to be the superior. However, in terms of tumor classification using the findings of the ROC graph, RF came out on top. This result was not in alignment with a similar study done by [16] where SVM and Random Forest Algorithm were used for the prediction of breast cancer. The sensitivity, classification rate as well as specificity was found to be higher with the RF algorithm when compared to SVM, concluding that the RF algorithm was more suited to extract suitable information. RF has also been known to be affected the least due to variance and overfitting and has a good scaling of the dataset which works in its favor as well [16]. However, a problem was encountered in preprocessing the image as it omitted some parts of the data. It not only reduced the image quality but also hindered the execution of the different ML algorithms.

B. Diabetes

The algorithm that is used for this purpose is known as Discriminant Analysis (DA). The input features are taken into consideration and several equations are executed on them. DA has a two-fold approach. The first

objective is the detection of an equation that can classify the test sample and the second one is to interpret it in a manner so that the relationship that exists between the features is properly identified [20]. Several factors can influence the classification process and some of them include the concentration of glucose in the body, the blood pressure, age of the patient, the thickness of the skin folds, and pregnancy.

The different algorithms that are used for this particular purpose are Gaussian Naive Bayes (GNB), CART, SVM, etc and they are further supplemented with data from medical records like Body Mass Index (BMI), serum glucose 1 and glucose 2 levels, gender, age, race and so on. They can be used to even predict Type 2 Diabetes accurately [21]. An important approach that was used here was the neural network algorithm. It involved a continuous interaction between a feedforward and a backpropagation algorithm. It also considered the features mentioned above, with the common ones being serum insulin levels, age, the thickness of the skinfold, pregnancies, etc. The neural networks were highly successful in making important predictions when compared to other algorithms employed for the same purpose [21]. The training of DNN and cross-validations has also been employed for the accurate detection of diabetes for a long time. Both methods have been stated to be highly accurate with a 97% prediction probability [22].

C. Kidney Diseases

Chronic Kidney Disease, also known as CDK, is one of the most common reasons for the failure of the functionality of the kidneys. The diagnosis of it is mainly based on the clinical manifestations, tests from the labs, biopsy, and imaging studies that are used for investigation. However, most of them have associated disadvantage. A biopsy is known to be risky and time-consuming as well as a costly procedure. The role of Machine Learning in a situation like this is important to overcome the disadvantages.

The most common classifier used here is the SVM, even though a lot of literature substantiating it is not present. Therefore, other classifiers like ANN and LR are also used for the same purpose. It was noticed that the ANN algorithm had the best results when compared to DT and LR for the early diagnosis of CKD [23]. The dataset for the Kidney Function Test (KFT) was used to compare the performance of different classifiers in the study. The classifiers used were KNN, RF, and NB, and the examination of the performance was done by checking the accuracy, precision, and F-measure [23]. In terms of both F-measure and accuracy, RF was found to be on the higher side while precision was found to be better in NB.

Another study conducted by [24] had a similar methodology and most of them concluded that SVM was the best classifier for early detection of kidney diseases. It could handle both unstructured and semi-structured data and also had a wide range of features for more accuracy.

In another study done by [25], the detection of kidney disease was done using the NB and the SVM classifier. The four diseases that were identified by the classifiers included Acute Nephritic Syndrome, Chronic Glomerulonephritis, CKD, and Acute Renal Failure. The features of comparison used in this particular study mainly the accuracy and the time of execution. In terms of accuracy, SVM scored higher than NB but the tables turned for execution time as NB was found to be better than SVM. Although all these studies pointed in favor of one parameter or the other, the suggestion that certain hyper-parameters were not taken into consideration weakened its claim. A study conducted by [4] concluded that the accuracy and performance results of the ML algorithms changed when the hyper-parameters were explored more comprehensively.

D. Lung Cancer

Lung Cancer is a common cause of death among males. It can start in the main airway of the lungs or the windpipe itself. It is seen more commonly in people with emphysema and cardiac problems. Lung Cancer related to small cell is the one that is the most difficult to detect as it closely resembles a physiological appearance [26]. Thus, Convolution Neural Network (CNN) has been employed for the premature detection of this cancer. Large datasets are generally preferred in this algorithm and this issue is solved using the Entropy Degradation Method (EDM) [27]. High-resolution CT scans are generally used for obtaining the data for testing as well as training. The analysis is manifested as histograms that are manifested and it is changed into scores which are further converted to logistic functions. One important aspect of this method is that the data that is obtained is either from a healthy person or a person suffering from SCLC. The initial test is performed by keeping this in mind. Although this method has been labeled to be accurate, it is still not considered to be the

best method for the detection of lung cancer. A larger dataset and using CNN for improving the image quality can be considered as the step in the right direction.

E. Acute Lymphoblastic Leukemia

Leukemia can be detected using microscopic images and the algorithms related to Machine Learning can be used to segment and classify them. The common algorithms that are used for leukemia are KNN, NB, and SVM with Multilayer Perceptron (MLP) are also being used from time to time. All of the algorithms work by dividing the data into sections of preprocessing, classification, extraction of features, and the final evaluation of the algorithm [28]. During preprocessing, the image is cropped so that only the area of interest is brought into focus and all the unnecessary information is removed. The noise of the picture is then removed to make it smoother and sharper, and it is done by the Gaussian Blur technique. Color, geometry, statistics, Haralick texture, binary patterns, and the presence or absence of cells adjacently are all adjusted in the stage of feature extraction.

F. Parkinson's Disease

[29] proposed a new framework for classification which was used for the diagnosis of PD. The algorithm was enhanced by a filter which improved the accuracy by 15%. Each subset of the dataset was taken, and classifiers were applied to it independently which compensated for any data loss. KNN, Discriminant Analysis, SVM, and NB were the classifiers that were used. SVM was found to be the most effective classifier in all aspects.

The fuzzy k-Nearest Neighbor (FKNN) algorithm was used by [30] as an effective means for the detection of Parkinson's disease. The study compared SVM and FKNN for the same purpose. The best FKNN model was constructed using Principal Component Analysis (PCA). The data that was taken from the same repository had voice recording with measurements from 31 people and 24 of them had Parkinson's disease. The study revealed that FKNN was more advantageous than SVM in almost all aspects. [31] went a step ahead by not only detecting but also tracking PD as it progressed, using multiple classifiers like Least Squares Support Vector (LS-SVM), GRNN, and MLPNN. This study also claimed that the SVM was the best performing model, and the result was further strengthened by decoding the metrics for optimal performance. Most of the authors propose frameworks in striking detail, as it is seen concerning kernel and regularization. An important point of consideration at this juncture is the calibration of the algorithms. It improves the classification process of various algorithms like NB, RF, and SVM effectively as proven by the study by [32].

G. Dermatological Disease

There is a lot of diversity in dermatological diseases, and it is very rare to find someone with expertise in it. Detection at an early stage ensures that there are no complications, and the outcomes are not serious. Some common dermatological diseases are psoriasis, eczema, and melanoma and a few of them can even be life-threatening.

The diagnosis of dermatological diseases using a Machine Learning algorithm starts with the step of collection and augmentation of data that are obtained from images. The second phase involves the creation and training of the model. The third and last model involves the conversion of the image into arrays which are broken down into pieces by the trained models. The augmentation can be done in several ways like Synthetic Minority Oversampling Technique (SMOTE) and other vision techniques like changing the color and contrast, smoothening, changing the greyscale, blurring and noise reduction, etc. As the size of the dataset increases, the appropriate training of the models becomes more important [33]. For example, training the CNN algorithm helps in addressing the problems related to overfitting. SVM takes input from various convolutional layers and it must be trained appropriately to do so. These SVM inputs get converted into vectors ultimately which is a form of storage as well.

H. Heart Diseases

The role of genetics in the etiology of diseases is a major factor of concern in precision medicine. Many areas of special interest in precision cardiology include oncology, genetics, and ischemic diseases. The most common methods for diagnosis in precision cardiology are like the ones used routinely and they include image and

blood tests more commonly or a combination of both. Since the root of several cardiovascular diseases has been related to its genetic predisposition, the role of different ML algorithms in precision medicines has gained a lot of importance in recent times [34]. The most common ones include Natural Language Processing (NLP), long short-term memory (LSTM), Recurrent Neural Network (RNN) and CNN with SVM also being commonly used. They can be utilized in deep learning methods for the accurate detection of heart diseases.

A study was done by [35] which took into consideration other parameters like Serum Cholesterol Levels and Resting Blood Pressure and Heart Rate to predict heart diseases even more precisely. 120 samples that were positive for heart diseases and 150 samples that were negative for the same were taken from the ML laboratory. The study compared the features of SVM, LR, NB, KNN, and various other classifiers. Accuracy and sensitivity were noted to be highest for LR classifiers which proved that it could be the most dependable for accurate diagnosis of heart diseases.

Another study was done by [36] which aimed at using the ML techniques for heart disease prediction. Different attributes like gender, chest pain, slope, and target were taken into consideration here. They deployed common algorithms like LR, DT, KNN, and NB. When accuracy was assessed, it was found to be highest in LR at 86.89%. [37] further reiterated the result using the Logistic Regression statistical test. Even the hyperparameters were taken into consideration in these studies which increased the credibility of the results of the ML algorithm. However, the fact that the datasets that were considered here to make this inference were relatively small could be a limiting factor.

I. Common Diseases

The use of ML algorithm for predicting the most common diseases was first proposed by [38]. The CNN and KNN techniques were used for the disease prediction based on the dataset available from the ML depository. CNN and KNN were compared by [38] for various features like accuracy and the time required for processing. CNN algorithm had an accuracy of 84.5% and a processing time of 11.1 seconds which stated it to be superior to KNN. The living habit and lifestyles of the patient were also taken into consideration to understand the risk factors associated with the disease being predicted. A similar result was shown in the study done by [39] which also harped on the superiority of CNN over the other algorithms such as Naive Bayes, KNN, and Decision Trees. The reason for the high accuracy of CNN was deemed to be the fact that it could detect nonlinear relationships of the most complex nature and provide a comprehensive description of the disease as well [38, 39]. These suggestions have sound scientific backing as well and are based on observations and statistical analysis. The only shortcoming related to it was the lack of details. Moreover, the feature of accuracy was given the highest importance which was not in line with the other prevalent theories [38]. The chances of bias concerning the different algorithms were also not taken into consideration [39].

Current Challenges

The chances of selection bias were avoided in this study by only picking up articles that had multiple ML algorithms that were supervised. Irrespective of the supervised algorithm being the same, the results that are obtained from it can be highly varied. Therefore, performing any comparison between them can create results that are not precise as the studies themselves are separate [4]. However, choosing studies with different variables can also cause a type of selection bias if the measurement for disease prediction is done using different algorithms.

The authors of the study did not consider all the variables that were available for a particular dataset. In fact, if a new variable would be included in the assessment, an underperforming algorithm could become more accurate, and it could be even more beneficial for prediction [40]. It is one of the most important limitations that the study encounters.

Other than the limitation mentioned above, the fact that the level of classification for the Machine Learning algorithm was broad, the comparison between various variables became a little cumbersome.

Another limitation that needs to be highlighted in the study is the non-involvement of the hyperparameters in the comparison for the different Machine Learning Algorithms that were chosen for consideration. An argument that has been regularly put forward is that if the hyperparameters are used, the same algorithm can have a varied result with contrasting accuracy even if the dataset remains the same [41]. Even if the selection of the kernels is different for the vector machines, the outcome accuracy can take a hit. It is applicable for a random forest model as well which generates varied results if the nodes are split.

Literature gap

[42] claims that the algorithm used in Machine Learning is prone to several errors mainly due to two reasons. First and foremost, the dataset that is selected should be of a high quality so that the results that are obtained do not have any bias and are precise. Secondly, the features that are selected for the extraction of information from the dataset should also be right. Overcoming these two issues became difficult due to several factors including the requirement of high power of computation. Any hindrance in this matter can be fatal to the lives of the patients. [43] was of a differing opinion as he believed that human errors by medical doctors could cause more risks in disease detection. Even though enough electronic medical data is available, doctors are unable to interpret it to detect diseases early. The ML algorithms can prove to be useful in this aspect; they can detect patterns and information that might otherwise remain hidden.

3. METHODS AND MATERIALS

Supervised Machine Learning Algorithm

Technically speaking, the algorithms that are associated with Supervised Machine Learning collect and analyze data to predict acceptable results. As more unique data is fed, the accuracy of the results also increases. The grouping of the Machine Learning algorithm can be done in several ways but the most accepted one is to group them into three categories with a broad range [44]. These categories are mainly in the form of unsupervised, semi-supervised, and supervised.

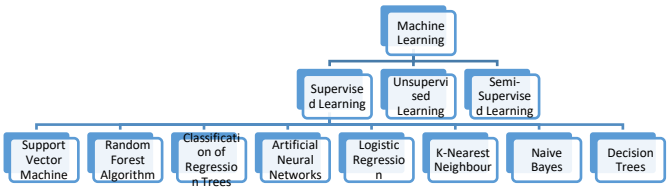

Figure 1: Categories of Machine Learning Algorithms [44]

The supervised algorithm involves the following steps. First, a previously labeled dataset is taken for the training of an algorithm. Once that is completed, the trained algorithm is executed on a dataset that is unlabeled so that the categorization can take place into similar groups [31]. For example, a dataset of three patients suffering from diabetes is taken in Figure 2. The figure shows how supervised machine learning algorithms perform for categorizing diabetic and non-diabetic patients based on abstract data.

Whenever there is a regression and a classification problem, supervised learning algorithms can be used. The output that is obtained from the classification problem is termed as discrete and it can be categorized independently as 'diabetic' or non-diabetics' or 'red' or 'black'. In the regression type, the output is of a more real representation like the risks of developing cardiovascular diseases.

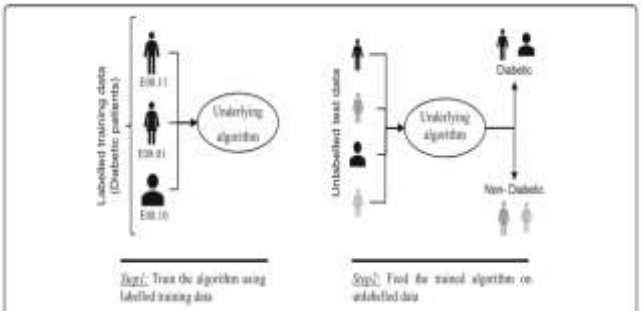

Figure 2: Representation of differentiation of the data into diabetic and non-diabetic patients [31]

The commonly used algorithms in Machine Learning have been discussed below.

1) Machine Learning Methods

a) Support Vector Machine (SVM)

It is another important supervised algorithm that mainly works for problems related to classification, but it is sometimes used for regression as well. The plotting of the data is initially done on a multidimensional space and the coordinates are based on the features used. The hyperplane concept is used to divide the data into separate classes. This leads to the maximization of the marginal distance between two hyperplanes taken for decision [45]. What sets SVM apart when compared to other algorithms is the ability to utilize nonlinear relationships accurately on map points. Since the data is divided into two separate classes, it is also known as a binary classifier of the non-probabilistic type. The accuracy of SVM is also higher than most other algorithms used for computations. However, the limitation of it is that it works only for smaller datasets. If larger datasets are taken, the training and computation become too complex for it to handle, and it takes up more time. Datasets with a lot of noise also act as a hindrance [4]. The creation of subsets for training is a technique to make classification using SVM better. Although it can be used for both linear as well as non-linear problems, it is the latter that gains more traction with the use of SVM. Fig. 3 provides the simplified illustration of an SVM classifier and shows the working of a support vector machine. A hyperplane (a line) has been identified by SVM which maximizes the separation between circles and star classes.

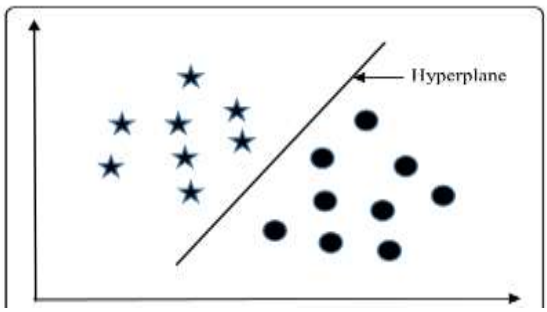

Figure 3: Illustration of SVM Classifier [45]

b) Random Forest Algorithm (RFA)

RFA is one of the most popular algorithms that can be used for both classification and regression problems [46]. The methodology that is primarily used in it is called recursion. It is based on the Decision Trees concept itself and the training is done using a method called bagging [46]. RFA is not affected by noise which makes it ideal for data that is imbalanced. Overfitting problems are also not a hindrance for RFA. Fig. 4 illustrates the RF Algorithm that consists of three different decision trees. Using a random subset of the training data, each decision tree was trained.

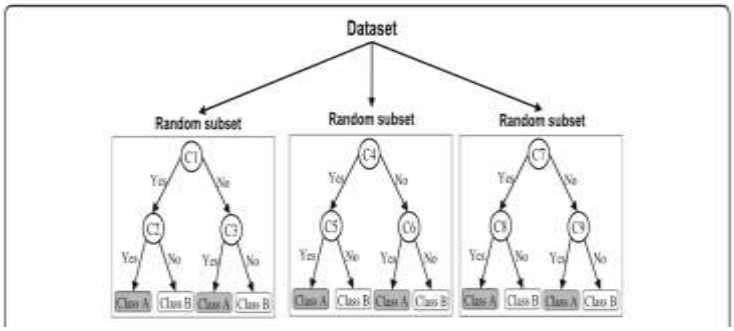

Figure 4: Illustration of RFA Classifier [46]

c) Classification of Regression Trees (CARTs)

It also resembles a tree model where the outcomes are deduced based on the values that already exist. The representation is primarily in the form of a binary tree and each input is represented by a root and then there is a split in the point in a variable [16]. The leaf node represents the output based on which the predictions are made.

d) Artificial Neural Networks (ANN)

It is a common Machine Leaning algorithm that is used for problems related to image classification. It is based on the concept of Artificial Neurons which is like what is observed biologically. It is a triple-layered model and each node in every layer is connected to other nodes present in the other layers [47]. As the layers get hidden, the creation of a neural network becomes more comprehensive. It is mainly based on three functions. Error function determines whether the output was good or bad and was it dependent on the input [48]. The search function determines the changes that can be done to rectify the common errors and the update function determines the implementation of these changes. It helps in the improvement of the performance of the algorithm. Fig. 5 depicts the illustration of ANN structure with two hidden layers with its interconnected group of nodes. The output of nodes from one layer is connected through arrows to the input of nodes of another layer.

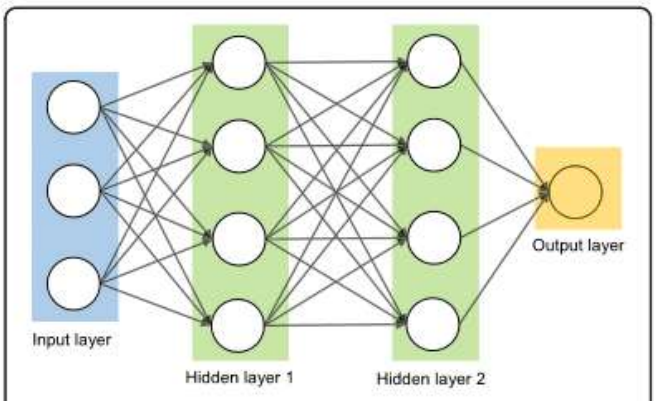

Figure 5: Illustration of ANN structure [47]

e) Logistic Regression

It is a procedure that is used to solve mathematical problems, using logistic functions, based on the datasets obtained from epidemiological studies. The coefficients of logistic regression are obtained, and the final predictions of the model are made after that [26]. The two parts of the model are linear and link function. Different calculations are executed using the linear portion while the output is delivered using the Link portion. A hypothesis and cost are two prime requirements of this supervised Machine Learning algorithm and good optimization of it helps in its proper execution [18].

f) K-Nearest Neighbour (KNN)

It is one of the most popular supervised algorithms that has its application in the detection of intrusion, recognition patterns, etc. The accuracy of KNN is relatively high even when the algorithm is simple to understand. The disadvantages associated with KNN are that it requires a high amount of energy as the testing as well as the training data have to be stored and the computation turns out to be pretty expensive [49]. Similar instances are first taken into consideration and then the output data is summarized to predict a result of the new inputs that are fed. It uses the mean of the values for regression while the mode is used for classification. Euclidean Distance measure is another important component that is utilized here. The vectors that are used for training should have been labeled separately. Fig. 6 shows a simplified illustration of the KNN algorithm. It shows that the sample object 'star' is classified as 'black' when K=3 as higher votes are from the 'black' class. However, the sample object is classified as 'red' when K=5 as higher votes are from the 'red' class.

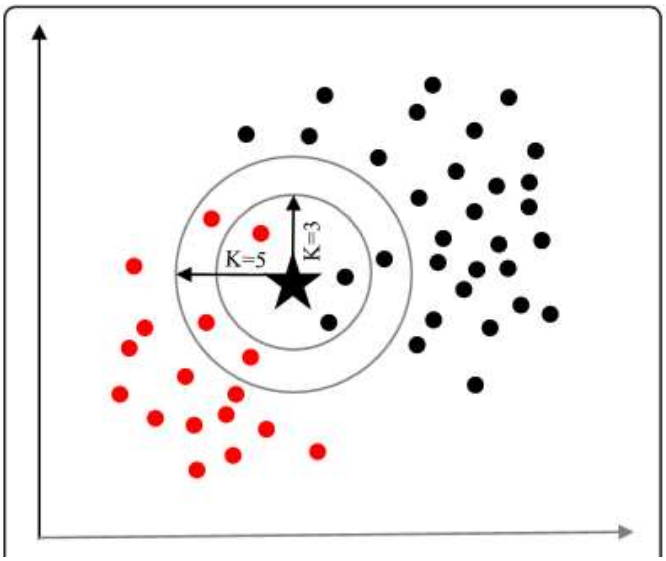

Figure 6: Illustration of KNN Algorithm [49]

g) Naive Bayes

Binary problems and the ones related to multiclass are resolved by this classification algorithm called Naive Bayes. It works on the principles of the Bayes Theorem. One common principle that is followed vehemently is that the features should be completely independent [16]. There is a striking similarity with SVM except that the statistical methods show some added advantage. As soon as new input is encountered, the probabilistic value is calculated immediately, based on the input [26]. The data with the highest value that is obtained from the input is also labeled. Fig. 7 shows an illustration of the Naive Bayes Algorithm with 'white' circle as the new sample that is required to be classified either to the 'green' or 'red' class.

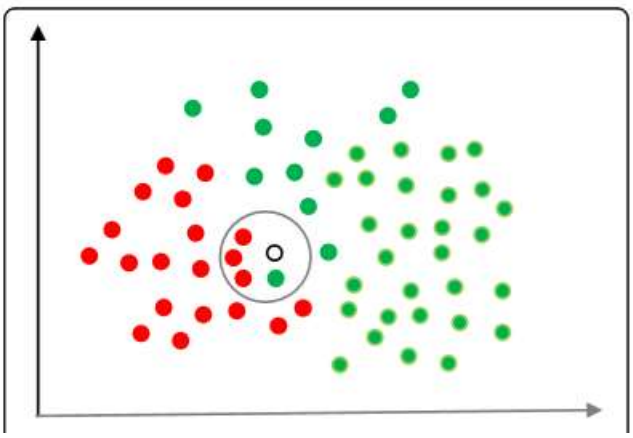

Figure 7: Illustration of Naive Bayes Algorithm [26]

h) Decision Trees (DT)

Three important things are considered in a DT algorithm, the decisions that are taken, the consequences that are possible, and the outcomes that can be generated from it. It has a tree-like representation with each node representing a question and the answers of it represented by branches. On reaching a leaf node, the sample will be assigned a corresponding node for consideration [44]. This is ideal for simple problems with small datasets. The problems of overfitting and bias in outcome are sometimes seen with DT, especially when imbalanced datasets are used. DT is also able to handle relationships of linear as well as non-linear types. Fig. 8 shows an illustration of the Decision Tree with its rules and elements depicted. The circle represents each variable which is C1, C2, and C3 and the rectangles depict the decision outcomes (Class A and Class B). Each branch is labeled with 'True' or False' depending on the outcome value from the test to its ancestor node to classify a sample to a class successfully.

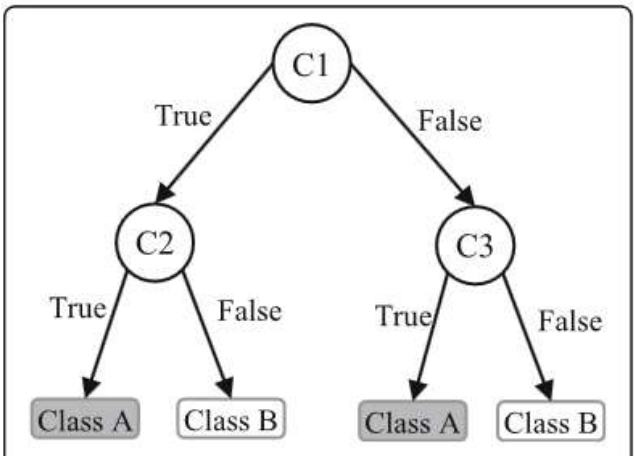

Figure 8: Illustration of Decision Tree with its rules and elements depicted [44]

Data source and Data extraction

The research was targeted at finding articles that had multiple supervised Machine Learning algorithms for the prediction of diseases. A search strategy was devised to include the appropriate articles. The following search terms were used for the search strategy-

- "Disease prediction" AND "machine learning"
- "Disease prediction AND "data mining"
- "Disease risk prediction" AND "machine learning"
- "Disease risk prediction" AND "machine data mining"

The study aimed to compare the supervised Machine Learning techniques of different types for the prediction of disease and the articles were selected based on this requirement. Python programming language was used to write an appropriate program that checked for the presence of multiple Supervised Algorithms in the title section, abstract, and also the Keyword section of the article.

4. RESULTS AND DISCUSSION

Results

The dataset that was obtained after conducting a thorough search of the literature had articles with more than one algorithm for the prediction of a single disease. The variants that were implemented have already been discussed in detail in the methodology section of this study. A generalized comparison was also made based on the measures of the performance metrics. The final articles were chosen based on these metrics, methodology, and the target disease.

The names and various references related to the disease as well as the Supervised Algorithms of Machine Learning to predict them have been included in Table 1. The algorithm that is most suited for a particular disease has also been discussed in the table.

Table 1: Applications of Machine Learning Algorithms in Medical Prediction

Disease Predicted	Algorithms compared	Prediction Performance (results)	Findings	Reference
Breast Disease	Bayesian Networks, RF, and SVM	Wisconsin Repository dataset and k-fold method of validation was used to detect breast cancer prematurely	Considering recall, accuracy, and precision, SVM was found to be the best	[19]
Breast Disease	SVM and Random Forest Algorithm (RFA)	RFA has dataset good scaling and got least affected due to overfitting and variance	RF algorithm was best suited among others	[16]
Diabetes	Gaussian Naive Bayes (GNB), CART, SVM, neural network	Used to predict Type 2 Diabetes	Neural networks made highly successful predictions	[21]
Diabetes	DNN and cross-validations	Accurate detection of diabetes was done	Both DNN and Cross-Validation were highly accurate with a 97% prediction probability	[22]
Kidney Diseases	SVM, ANN, and LR	Early diagnosis of Chronic Kidney Disease (CKD)	ANN had the best results compared to DT and LR	[23]
Kidney Diseases	KNN, RF, and NB	Kidney Function Test (KFT) was done to evaluate performance by checking the accuracy, precision, and F-measure	RF was found to be higher in F-measure and accuracy, while precision was found to be better in NB	[23]
Kidney Diseases	SVM	SVM can handle both semi-structured and unstructured data and has the capacity of giving more accuracy	SVM is the best classifier	[24]
Kidney Diseases	NB and SVM	Features were used to detect Acute Nephritic Syndrome, Chronic Glomerulonephritis, CKD, and Acute Renal Failure	In terms of accuracy, SVM scored higher than NB but the NB was found to be better than SVM in execution time	[25]
Lung Cancer	Convolution Neural Network (CNN)	Entropy Degradation Method for(EDM) is used premature detection of	High-resolution CT Scans are used for obtaining the	[27]

		this particular type of cancer	data testing and data training	
Acute Lymphoblastic Leukemia	KNN, NB, and SVM and Multilayer Perceptron (MLP)	Data is divided into sections of preprocessing, classification, extraction of features, and the final evaluation of the algorithm	All algorithms performed better	[28]
Parkinson's Disease	KNN, Discriminant Analysis, SVM, and NB	Accuracy was improved by 15%.	SVM was found to be the most effective classifier in all aspects	[29]
Parkinson's Disease	SVM and FKNN	The data that was taken from the same repository had voice recording with measurements from 31 people and 24 of them had Parkinson's disease	FKNN was better than SVM in almost all aspects	[30]
Parkinson's Disease	Least Squares Support Vector (LS-SVM), GRNN, and MLPNN	Used for detecting and tracking the disease	SVM was the best performing model	[31]
Dermatological Disease	CNN, SVM	SVM takes input from various convolutional layers Training the CNN algorithm helps in addressing the problems related to overfitting.	SVM and CNN both performed better	[33]
Heart Diseases	SVM, LR, NB, KNN	Serum Cholesterol Levels and Resting Blood Pressure and Heart Rate were evaluated to predict heart diseases	LR classifiers performed highly in terms of sensitivity and accuracy	[35]
Heart Diseases	LR, DT, KNN, and NB	Highest Accuracy of 86.89% was found with LR	LR performed better than others	[36]
Common Diseases	CNN and KNN	CNN algorithm had an accuracy of 84.5% and a processing time of 11.1 seconds	CNN is superior to KNN	[38]
Common Diseases	CNN, Naive Bayes, KNN, and Decision Trees	High accuracy of CNN was found to detect nonlinear relationships of the most complex nature of the disease	CNN was found to be superior	[39]

Discussion

The Machine Learning Algorithm performance is something that needs to be discussed. The accuracy and log loss are two factors that are instrumental in judging it. For an algorithm to be well-performing, the accuracy should be as high as possible while the log loss should be very low. Whenever a choice needs to be made regarding the most suitable algorithm among all the available options, these factors should take the first precedence.

All the papers that are included in this study and have been referenced concerning various diseases have been published between the years 2015 to 2021.

Declaring an algorithm to be the best or even better than any other algorithm is not possible because the domain in which it is tested and trained does not remain constant. Certain other factors like the dataset that is used for training as well as the testing, the features that are used on the algorithm, the pre and post-processing, the type of the dataset in general, the size of it, the machine that is used to analyze, the level of performance, the machine capacity, all of them also play a role in establishing a comparison. These all are instrumental in selecting the right algorithm. When there is a problem at hand, the selection process does not take place immediately. In fact, it is a gradual process consisting of multiple reiterations which filter through the available set of algorithms. The creation of the filtered set is done after going through past experiences and gaining vast knowledge. Its application in the medical field is based on the fact that the diagnosis of a disease can be done prematurely using the correct set of algorithms. The scenarios given below can provide an even better perspective of the situation.

The diagnosis of kidney disease has been mainly done using ANN, LR, as well as DT algorithms but ANN was found to be most suitable in this regard by a large margin [23]. The result is a little different when seen in terms of lung diseases as SVM outperforms most of the other algorithms like ANN and DT. One major reason for the success of SVM is the fact that the stage of cancer of the lungs can also be detected by SVM [50]. Regression is a statistical tool to establish an association between the dependent and the independent variables. They are of two major types, linear and logistic regression. In linear regression, the dependent variable is of a continuous type while it is of discrete type for logistic regression. Serious diseases like cancer should not be predicted using logistic regression. KNN also cannot be used for cancer as it results in clustering due to the large datasets that are used for its prediction [51]. DNN performs better for predicting breast cancer. As far as performance was concerned, ANN took the top spot, with SVM and KNN coming second and third respectively [52]. DNN has also been found to be accurate for the detection of diabetes when compared to other supervised algorithms related to Machine Learning.

5. CONCLUSION

The integration of the field of Machine Learning with the medical field has gained a lot of prominence in recent times as it helps in the accurate and early prediction of disease using the available data. This study has given a brief overview of the most important Supervised Machine Learning algorithm which plays a crucial role in identifying the diseases of the kidney, heart, liver, breast cancer, diabetes, etc. The result obtained by the researchers has also been provided in a tabulated format.

Different algorithms were employed to identify the problems related to the heart, brain, kidney beforehand. According to a thorough review of literature, the algorithms that were widely used for prediction were SVM, LR, and RF and they were found to be more accurate in terms of performance as well. The CNN model was used for the most accurate prediction of the diseases that are commonly seen. The SVM model was found to be superior in the case of Parkinson's disease and chronic kidney disease as the data obtained for its analysis was either of an unstructured or semi-structured type. Since the detection of breast cancer needs a proper classification first, the RF algorithm was found to be most suitable for it as it could deal with a larger set of data and did not get affected by overfitting. Finally, the LR algorithm was used for detecting heart diseases reliably.

It was also found that the accuracy of the algorithms was not consistent, and it changed with the different datasets themselves. It depended on the selection of features, the number of them, the performance metrics of the model, etc. It was also surmised in the study that the accuracy, as well as the performance of the model, can be made better by creating an ensemble that uses multiple algorithms at once.

Future Work

Looking at the problems encountered in the studies that were included in this paper, it has become evident that more complex algorithms concerning Machine Learning need to be created for accurate and efficient prediction of diseases. Moreover, the training models must be calibrated at regular models so that the performance improves. There should also be an expansion of the demographic details so that the problems related to overfitting are not encountered. Lastly, the overall performance of the model can be improved by making the correct feature selection for analysis.

Open Research Questions

The application of Machine Learning in various sections of the given field is commendable. It can be used in the diagnosis as well as the prediction of diseases using imaging and event extraction that play a very crucial role in the field of healthcare. One thing that has become increasingly evident is that there has been a complete integration of computational biology with machine learning, and it is not in its nascent stage anymore. It has crossed most of the major hurdles that were present on its path and has reached its peak. The introduction of the concept of precision medicine is further testimony of it.

The methodology that has been proposed here can help to improve the prediction accuracy of a disease so that it can be identified at an earlier stage. This can affect the treatment plan as well as it can be devised accordingly. The estimation of the accuracy of the different ML algorithms can be done to enhance its effect in the future. The model is not completely free of limitations though. First, since the data being dealt with is of a higher magnitude, the processing time will be more, especially when it is used for the training data. The algorithms can be implemented on data collected in real-time in the future as well as it will determine whether the system is effective or not. Looking at all the limitations, the need to have a complex combination of algorithms is imperative for creating a model with high accuracy and early prediction.

6. REFERENCES

[1] K. Mtonga, S. Kumaran, C. Mikeka, K. Jayavel and J. Nsenga, "Machine Learning-Based Patient Load Prediction and IoT Integrated Intelligent Patient Transfer Systems," *Future Internet,* vol. 11, no. 236, 2019.

[2] T. Davenport and R. Kalakota, "The potential for artificial intelligence in healthcare," *Future Healthcare Journal,* vol. 6, no. 2, p. 94–98, 2019.

[3] A. Gavhane, G. Kokkula, I. Pandya and K. Devadkar, "Prediction of heart disease using machine learning," in *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 2018.

[4] S. Uddin, A. Khan, M. Hossain and M. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC Medical Informatics and Decision Making,* vol. 19, no. 1, pp. 1-16, 2019.

[5] R. Katarya and P. Srinivas, "Predicting heart disease at early stages using machine learning: A survey," in *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, 2020.

[6] Sidey-Gibbons, A. M. Jenni and C. J. Sidey-Gibbons, "Machine learning in medicine: a practical introduction," *BMC Medical Research Methodology,* vol. 19, 2019.

[7] D. Singh and V. Kumar, "Single image defogging by gain gradient image filter," *Science China Information Sciences,* vol. 62, no. 7, pp. 1-3, 2019.

[8] S. Otoum, B. Kantarci and H. Mouftah, "On the feasibility of deep learning in sensor network intrusion detection," *IEEE Networking Letters,* vol. 1, no. 2, p. 68–71, 2019.

[9] R. Razavi-Far, E. Hallaji and M. Farajzadeh-Zanjani, "Information fusion and semi-supervised deep learning scheme for diagnosing gear faults in induction machine systems," *IEEE Transactions on Industrial Electronics,* vol. 66, no. 8, p. 6331–6342, 2019.

[10] G. Shaheamlung, H. Kaur and M. Kaur, "Survey on machine learning techniques for the diagnosis of liver disease," in *International Conference on Intelligent Engineering and Managemen*, 2020.

[11] D. Q. Zeebaree, A. Abdulazeez, D. A. Zebari, H. Haron and H. N. Abdull Hamed, "Multi-Level Fusion in Ultrasound for Cancer Detection Based on Uniform LBP Features," *Computers, Materials & Continua,* vol. 66, no. 3, p. 3363–3382, 2021.

[12] F. K. Alsheref and W. H. Gomaa, "Blood Diseases Detection using Classical Machine Learning Algorithms," *International Journal of Advanced Computer Science and Applications,* vol. 10, no. 7, 2019.

[13] T. Untawale, "A REVIEW ON MACHINE LEARNING TECHNIQUES TO PREDICT DISEASES," *International Research Journal of Modernization in Engineering Technology and Science,* vol. 2, no. 7, 2020.

[14] S. Katiyar and S. Jain, "Predictive Analysis on Diabetes, Liver and Kidney Diseases using Machine Learning," *International Journal for Research in Applied Science & Engineering Technology (I.J.R.A.S.E.T.),* vol. 8, no. 5, 2020.

[15] M. Fatima and M. Pasha, "Survey of Machine Learning Algorithms for Disease Diagnostic," *Journal of Intelligent Learning Systems and Applications,* 2017.

[16] A. Bharat, N. Pooja and R. Reddy, "Using machine learning algorithms for breast cancer risk prediction and diagnosis," in *Proceedings of the 3rd International Conference on Circuits, Control, Communication and Computing*, Bangalore, India, 2018.

[17] H. Dhahri, E. Al Maghayreh, A. Mahmood, W. Elkilani and M. Faisal Nagi, "Automated breast cancer diagnosis based on machine learning algorithms," *Journal of Healthcare Engineering,* pp. 1-11, 2019.

[18] M. Al-Hadidi, A. Alarabeyyat and M. Alhanahnah, "Breast cancer detection using K-nearest neighbor machine learning algorithm," in *Proceedings of the 9th International Conference on Developments in eSystems Engineering*, Liverpool, UK, 2016.

[19] P. G. M. Sengar and A. Nagdive, "Comparative study of machine learning algorithms for breast cancer prediction," in *Proceedings of the 3rd International Conference on Smart Systems and Inventive Technology, ICSSIT 2020*, 2020.

[20] A. Al-Zebari and A. Sengur, "Performance comparison of machine learning techniques on diabetes disease detection," in *Proceedings of the 1st International Informatics and Software Engineering Conference*, Ankara, Turkey, 2019.

[21] S. M. D. A. Chinthaka Jayatilake and G. U. Ganegoda, "Involvement of Machine Learning Tools in Healthcare Decision Making," *Journal of Healthcare Engineering,* pp. 1-20, 2021.

[22] S. Ayon and M. Islam, "Diabetes prediction: a deep learning approach," *International Journal of Information Engineering and Electronic Business,* vol. 7, no. 6, p. 21–27, 2019.

[23] Y. Amirgaliyev, S. Shamiluulu and A. Serek, "Analysis of chronic kidney disease dataset by applying machine learning methods," in *Proceedings of the IEEE 12th International Conference on Application of Information and Communication Technologies*, Almaty, Kazakhstan, 2018.

[24] P. Kotturu, V. Sasank, G. Supriya, C. Manoj and M. Mahesh- warredy, "Prediction of chronic kidney disease using machine learning techniques," *International Journal of Advanced Science and Technology,* vol. 28, no. 16, p. 1436–1443, 2019.

[25] A. Charleonnan, T. Fufaung, T. Niyomwong, W. Chokchueypattanakit, S. Suwannawach and N. Ninchawee, "Predictive analytics for chronic kidney disease using machine learning techniques," in *2016 Management and Innovation Technology International Conference, MITiCON 2016*, 2017.

[26] P. Radhika, R. Nair and G. Veena, "A comparative study of lung cancer detection using machine learning algorithms," in *Proceedings of the IEEEInternational Conference on Electrical, Computer and Communication Technologies,*, Coimbatore, India, 2019.

[27] Q. Wu and W. Zhao, "Small-cell lung cancer detection using a supervised machine learning algorithm," in *International Symposium on Computer Science and Intelligent Controls*, Budapest, Hungary, 2017.

[28] S. Mandal and V. Daivajna, "Machine learning based system for automatic detection of leukemia cancer cell," in *Proceedings of the IEEE 16th India Council International Conference*, Rajkot, India, 2019.

[29] M. Behroozi and A. Sami, "A multiple-classifier framework for Parkin- son's disease detection based on various vocal tests," *International Journal of Telemedicine and Applications,* 2016.

[30] N. K. Dastjerd, O. C. Sert, T. Ozyer and R. Alhajj, "Fuzzy classification methods based diagnosis of Parkinson's disease from speech test cases," *Curr. Aging Sci.,* vol. 12, p. 100–120, 2019.

[31] S. Aich, H. Kim, K. Younga, K. Hui, A. Al-Absi and M. Sain, "A supervised machine learning approach using different feature selection techniques on voice datasets for prediction of Parkinson's disease," in *Proceedings of the 21st International Conference on Advanced Communication Technology*, PyeongChang Kwangwoon_Do, Korea (South), 2019.

[32] M. F. Ferjani, Disease Prediction Using Machine Learning, Bournemouth, England: Bournemouth University, 2020.

[33] Y. Hasija, N. Garg and S. Sourav, "Automated detection of dermatological disorders through image-processing and machine learning," in *Proceedings of the International Conference on Intelligent Sustainable Systems*, Palladam, India, 2017.

[34] S. Niazi, H. Khattak, Z. Ameer, M. Afzal and W. Khan, "Cardiovascular care in the era of machine learning enabled personalized medicine," in *Proceedings of the International Conference on Information Networking*, Barcelona, Spain, 2020.

[35] A. K. Dwivedi, "Performance evaluation of different machine learning techniques for prediction of heart disease," *Neural Computing and Applications,* vol. 29, no. 10, p. 685–693, 2018.

[36] M. Marimuthu, M. Abinaya, K. Madhankumar and V. Pavithra, " Review on Heart Disease Prediction using Machine Learning and Data Analytics Approach," *International Journal of Computer Applications,* vol. 181, no. 18, p. 20–25, 2018.

[37] K. Polaraju, D. Durga Prasad and M. Tech Scholar, "Prediction of Heart Disease using Multiple Linear Regression Model," *International Journal of Engineering Development and Research,* vol. 5, no. 4, p. 2321–9939, 2017.

[38] D. Dahiwade, G. Patle and E. Meshram, "Designing disease prediction model using machine learning approach," in *Proceedings of the 3rd Inter- national Conference on Computing Methodologies and Communication ICCMC 2019*, 2019.

[39] S. Jadhav, R. Kasar, N. Lade, M. Patil and S. Kolte, "Disease Prediction by Machine Learning from Healthcare Communities," *International Journal of Scientific Research in Science and Technology,* p. 29–35, 2019.

[40] O. Levy, Y. Goldberg and I. Dagan, "Improving distributional similarity with lessons learned from word embeddings. Tra," *Trans Assoc Comput Linguistics,* vol. 3, pp. 211-225, 2015.

[41] M. Lucic, K. Kurach, M. Michalski, O. Bousquet and S. Gelly, "Are GANs created equal? a large-scale study," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018.

[42] F. Yuan, "Critical issues of applying machine learning to condition monitoring for failure diagnosis," in *2016 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, 2016.

[43] S. Ismaeel, A. Miri and D. Chourishi, "Using the extreme learning ma- chine (elm) technique for heart disease diagnosis," in *2015 IEEE Canada International Humanitarian Technology Conference (IHTC2015)*, 2015.

[44] D. Kumar, J. P. Kumar, V. Prakash and K. Divya, "Supervised Learning Algorithms: A Comparison," *Kristu Jayanti Journal of Computational Sciences,* vol. 1, no. 1, pp. 1-12, 2020.

[45] A. M. Abdulazeez, M. A. Sulaiman and D. Q. Zeebaree, "Evaluating Data Mining Classification Methods Performance in Internet of Things Applications," *Journal Of Soft Computing And Data Mining,* vol. 1, no. 2, pp. 11-25, 2020.

[46] M. Yarabarla, L. Ravi and A. Sivasangari, "Breast cancer prediction via machine learning," in *Proceedings of the 3rd International Conference on Trends in Electronics and Informatics*, Tirunelveli, India, 2019.

[47] M. Ahmed, S. Hasan Mahmud, M. Hossin, H. Jahan and S. Haider Noori, "A cloud based four-tier architecture for early detection of heart disease with machine learning algorithms," in *Proceedings of the IEEE 4th International Conference on Computer and Communications*, Chengdu, China, 2018.

[48] B. Erickson, P. Korfiatis, Z. Akkus and T. Kline, "Machine learning for medical imaging," *RadioGraphics,* vol. 37, no. 2, pp. 505-515, 2017.

[49] W. Hussain and O. Sohaib, " Analysing Cloud QoS Prediction Approaches and Its Control Parameters: Considering Overall Accuracy," *IEEE Access,* vol. 7, p. 82649–8267, 2019.

[50] W. Rahane, H. Dalvi, Y. Magar, A. Kalane and S. Jondhale, "Lung cancer detection using image processing and machine learning healthcare," in *Proceedings of the International Conference on Current Trends towards Converging Technologies*, Coimbatore, India, 2018.

[51] S. Saxena and S. Prasad, "Machine learning based sensitivity analysis for the applications in the prediction and detection of cancer disease," in *Proceedings of the IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and RoboticsProceedings of the IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics*, Manipal, India, 2019.

[52] R. Chtihrakkannan, P. Kavitha, T. Mangayarkarasi and R. Karthikeyan, "Breast cancer detection using machine learning," *International Journal of Innovative Technology and Exploring Engineering,* vol. 8, no. 11, pp. 3123-3126, 2019.

INFO

Corresponding Author: ***Chetanpal Singh, Holmes Institute, Melbourne Australia.***

How to cite this article: ***Chetanpal Singh, Machine Learning Applications Related to Medical Predictions, Asian. Jour. Social. Scie. Mgmt. Tech.2022; 4(2): 127-144***