Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 1, 2021, Pages. 3495 - 3511
Received 15 December 2020; Accepted 05 January 2021.

Chronic Diseases Prediction Using Machine Learning – A Review

Shweta Agarwal[1], Dr.Chander Prabha[2], Dr.Meenu Gupta[3]

[1]PhD Research Scholar, Department of Computer Science and Engineering, Chandigarh University, Gharuan, Punjab, India
[2]Associate Professor, Department of Computer Science and Engineering, Chandigarh University, Gharuan, Punjab, India
[3] Assistant Professor, Chandigarh University, Punjab
Email: ershweta.cs@gmail.com[1], prabhanice@gmail.com[2]
shweta.e9500@cumail.in[1], chander.e9251@cumail.in[2], menu.e9406@cumail.in[3]

Abstract
The only way to overcome with the mortality due to chronic diseases is to predict it earlier so that the disease prevention can be done. Such model is a Patient's need in which Machine Learning is highly recommendable.The main objective is to collect all published articles related to disease prediction and conclude about the coverage of research done so far.For conducting this survey, we focused on published research from 2017 up to today, review the research done on various machine learning algorithms used for the efficient prediction of diseases.It has been observed that features, independent variable selection and combination of multiple algorithms plays an important role in improving the accuracy as well as performance of a disease prediction system, and it is possible to diagnose people based on symptoms.In this paper, we discussed different ML techniques and their accuracy that other researchers used to diagnose chronic diseases.

Keywords: Healthcare, Patients' needs, Disease prevention, Applications, Public health, Machine Learning, Disease Prediction, Chronic diseases

INTRODUCTION

Now-a-days, people are facing many health-related problems due to their living habits and environmental condition. Every single person amongst four is suffering from some chronic disease. Chronic diseases (such as cancer, diabetes and chronic diseases) are the major cause of death globally. These are the diseases that persevere for a long time, and are difficult to cure but disease prevention can only be possible if these diseases can be predicted earlier.

In 2010, as per the United States data, 67.2% of mortality in females and 65.8% of mortality in males[1]were due to these chronic diseases. However in India, amongst 10·3 million mortality happened in the year 2004, nearly 1·1 million (11%) people die due to injuries and 5·2 million (50%) people die due to chronic diseases[2]. In comparison to high-income countries, death rate due to chronic diseases for age-specific people is higher in India. Approximately 61% of death cause only due to Cancer, Diabetes and heart diseases, they record 55% of the mortality in the age gathering of 30-69 years [3] and it will be increasing coming years.

Public health is a major issue today, the lifestyle that we have achieved has rewarded us so many diseases, and people spent large amount on the treatment of these diseases[4]. The only way to overcome with this mortality rate and to prepare for the future disease [5] is to predict it at an early stage so that proper treatment can be provided to the patient, and it can be prevented at an early stage. Such disease prediction model is a Patient's need, which is feasible, robust, accurate and reliable to predict diseases on time [6].

To reduce the risk of disease and to detect it at an earlier stage, there is one IoT based model [7], but the main disadvantage of this model is that patient needs to wear more devices which is not comfortable to carry for the patient [7]. Therefore, a prediction model is required that does not require any device to carry but can predict the disease at initial stage by just taking the information or medical history of a patient. The data can be collected through different sources viz, interviews, observations, imaging reports, laboratory tests, therapies, treatment, bills, surveys and insurance[8]. The data acquired from these various healthcare sources are the information related to patients, diseases, hospitals and so on. It is required to extract useful information from the data collected because laboratory reports may contain some degree of error and patient could fail to describe their symptoms. This causes trouble for a doctor to predict accurate diseases due to which disease prevention is a big challenge[9]. Hence, this can be realized and analysed using machine learning algorithms to provide information, which will be a great help for healthcare system.

Machine Learning

Machine learning is highly recommendable and widely used in various fields like Security, finance, healthcare etc. Machine Learning: the classic definition is - A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E [10]. It is a type of AI that provides systems with the ability to learn and develop automatically from experience without being explicitly programmed [11]. Machine leaning is how computers recognize patterns and make decisions without being explicitly programmed. With ML, instead of programming a computer step-by-step, we can program a computer to learn just like we learn, through trial and error, and lots of practice.

2.1. Role of ML in Prediction

Machine Learning learns from experience. Here, "experience" is "lots of data". It can take any kind of data – images, video, audio or text and begin to recognize pattern in that data. With the help of machine learning technique, machines learn how to handle the data more efficiently as we cannot extract the meaningful information from data by just viewing it, but machine learning can do [12]. Its purpose is to learn from data [13]. A lot of researches have been done on how machine learn by themselves [14][15].

Once it learns to recognize patterns in data, it can also learn to make predictions based on those patterns. Machine learning has a potential to substantially improve prediction which is often used in conjunction with large data sets. It consists of so many efficient algorithms, frameworks & applications to achieve greater correctness of prediction.

Received 15 December 2020; Accepted 05 January 2021.

Stages to apply Machine Learning on data

The role of machine learning can be divided into seven main stages [16] which are shown in Figure 1.

(i) **Data Gathering:** Either the data is written on paper, documented in text files and spreadsheets, or stored in a database system, there is always a requirement to process it in an electronic format so that it can be suitable for analysis. This step is very critical because the quality and amount of data that we obtain, directly influences the quality of ML project. This data will serve as the learning material used by an algorithm to produce actionable knowledge.

(ii) **Data Preparation:** Once the data is gathered, then it is loaded into a suitable place and prepare it for use in ML training.

(iii) **Choosing a Model:** Data scientists and researchers have already created so many models. These models are suited for sequences (viz text or music), image data, text-based data and numerical data. An appropriate algorithm will be selected and the data in the form of a model be represented.

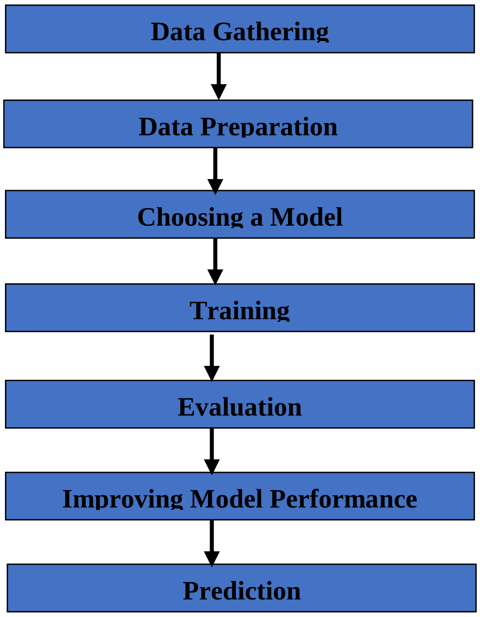

Figure 1: Seven stages for the implementation of Machine Learning on data

(iv) **Training:** By the time model is chosen and data has been prepared for analysis, then it will be used to steadily improve the ability of a model to predict the results accurately.

(v) **Evaluation:** As ML models lead to a biased solution to the problem of learning, it is therefore important to determine how well the algorithm has learned from its experience. This can be done by testing the model against that data which has not been used for training. This helps us to see how the model will work against data that has not yet seen. Training – evaluation is usually divided in the range of 70%-30% or 80%-20%.

(vi) **Improving model performance:** Once the evaluation is done, it is possible that we want to further improve the training. This can be done by tuning some of the parameters. Hence to improve model performance, testing is done on different assumptions.

(vii) **Prediction:** After all the above steps have been completed, Prediction is the last step to do something useful by deploying the model. ML uses data to answer questions, so prediction is the stage where some questions can finally be answered. This is the point where the value of ML is realized in all of this work.

Machine Learning Algorithms

Figure 2: Machine Learning Types

(i) Supervised Learning: It is that kind of ML algorithm which uses a known dataset also referred to as training dataset to form classifications or predictions [17]. This dataset includes labelled data that consist of input data and response values. For example, solutions are provided together with every problem.

(ii) **Unsupervised Learning:** It is that kind of ML algorithm which is used to draw inferences from data sets consisting of input data without labelled responses[18]. For instance, when kids start taking decisions out of their own understanding.

(iii) **Reinforcement Learning:** It is that kind of Machine Learning which learns from its own experience[19]. For example, if a new situation comes up, kid will take actions on their own, from the past experience, but parent can tell whether the action is good or not.

Apart from all above three main algorithms, so many algorithms are also there that are the part of these three ML algorithm types. A brief overview of all other algorithms along with their uses in different applications is mentioned in table below –

Table 1.Overview of different algorithms and its applications

Types of Learning	Model / Method	Various extended algorithms	Applications
Supervised Learning	Classification	Naive-Bayes, k-nearest neighbor, logistic regression, decision tree, random-forest, SVM	Fraud Detection, Email Spam Detection, Diagnostics, Image Classification

	Regression	Linear Regression, Polynomial Regression, Regression Trees, Bayesian Linear Regression, Non-Linear Regression	Risk Assessment, Score Prediction
Unsupervised Learning	Dimentionality Reduction	-	Text Mining, Face Recognition, Big data Visualization, Image Recognition,
	Clustering	K-means clustering, hierarchical clustering	Biology, City Planning, Targeted Planning,
	Association	-	-
Reinforcement Learning	-	-	Gaming, Finance Sector, Manufacturing, Inventory, Management, Robot Navigation, Trading, Online Recommendation, Text Mining

Applications of Machine Learning in Real Life

Let’s take a look on Machine Learning applications see how this technology can be applied to real-life problems-

Figure 3: Applications of Machine Learning

(i) **Image Recognition**: It is the most significant application of Machine Learning. It is an approach for detecting and identifying an object, place, person or the feature in a digital image. Various examples of this technique are – face detection, pattern recognition, face recognition, OCR and many more. The use of ML is to analyze the image pixel by pixel and extract the features of an image. Facebook provides.

(ii) **Speech Recognition (SR):** Machine learning helps the software to adapt to dynamic speech patterns users use idioms slang, abbreviations and to stay flexible. This is where machine learning is essential. Even theoretically, a human team can't teach millions of speech variations to the software manually. If the system trains itself, however the task becomes much more manageable.

(iii) **Healthcare Services:** ML methods brings an innovation in healthcare industry. It is extensively used in medical problems to disease prediction and diagnosis, medical research, therapy, support and planning. According to researchers, ML plays an important role in accurate identification of disease that help to facilitate medical experts so that the quality of medical care can be improved. The measurement in ML applications are the results of medical diagnosis such as different medical images, medical tests viz. blood test, blood pressure test etc., presence or absence of various symptoms and general information of patient like age, weight etc. On the basis of these results of measurement, doctors narrow down on the disease inflicting the patient.

(iv) **Prediction:** Prediction is the process of determining something based on previous history. It can be house price prediction, weather prediction, disease prediction, traffic prediction, and many more. Every kind of forecast is possible with ML approach. There are so many algorithms used to accomplish this task. These algorithms are useful in predicting and diagnosing chronic diseases in healthcare [20].

(v) **Sentiment Analysis:** The main task of Sentiment Analysis is to predict "what other people think?". For example, someone has written that "the product is good", then to find out the exact opinion or thought of that person from a text that "is it good or not". It is applied on decision making applications, review based website. The role of ML is to extract the knowledge from data by using both supervised and unsupervised learning.

(vi) **News Classification:** As the amount of information is growing exponentially, individual user need tools that would classify and sort the information as per their interest and choice. Hence ML algorithms can be useful to run through millions of articles in many languages and select the ones that are relevant to user interests and habits.

(vii) **Video Surveillance:** Machine learning can help develop complex algorithms for video recognition at first using human supervision. The system can help to spot any suspicious objects, human figures, unknown cars etc. Soon it will be possible to imagine a video surveillance system that functions entirely without human supervision.

(viii) **Email Analysis:** Machine algorithm can analyze and compare legitimate emails with spam and determine differences even in cases where humans would easily make a mistake.

(ix) **Cyber Security:** ML algorithms can immediately detect cyber security threats. The system will recognized the threat analysed similar cases and take measure to secure the website or application. It allows businesses to be up-to-date with malicious practices and predict safety issues before they even come up.

(x) **Author Identification:** As we know that the use of Internet is tremendously growing, due to which its illegal use for inappropriate purposes is a major concern now-a-days. Thus, ML algorithm helps in author identification so that crimes can be stopped.

(xi) **Social Media Services:** Social media like Facebook and many other are using techniques of Machine Learning to continuously monitor our activities and based on the activities monitored provide us so many attractive features like – suggestion to comment, whom to chat, friends suggestions etc.

(xii) **Recommendation System:** This is one of the advance applications of ML. Many search engines like Google and online shopping websites are using this feature in which similar type of websites, products and services are recommender to a user after a purchase or search. Various Machine Learning algorithms are used to implement recommendation system.

(xiii) **Online Customer Service:** To provide online customer support, websites develop a chat BOT as a representative to handle customer queries. This can be done with the help of Machine Learning algorithms; it analyses customer behaviour from the data chat. Bot developers can know which issues to focus on. As soon as several dozens of responses were confirmed, the chat BOTS can learn on their own from daily interactions with clients getting better with each dialog.

(xiv) **Age/Gender Identification:** Machine Learning is so advanced that it can help to identify age as well as gender of a person by using its one of the algorithm that is SVM classifier. This application is very much useful in forensic task.

(xv) **Language Identification:** Machine Learning is most efficient approach in identifying the type of a Language. Apache Tika, Apache Open NLP are the most common software used for language identification.

(xvi) **Information Retrieval:** As we know that data is growing tremendously on the web, hence IR plays a major role in big data. It is the method of extracting from unstructured data meaningful information. ML uses user habits and interests from analyzing search statistics and provides the result. Now, rating algorithms won't rely on Meta tags and keywords but instead will analyze the context of the page. Google RankBrain is the great example of this idea.

(xvii) **Robot Control:** To gain control over the features of drone, helicopter, robot etc, machine learning algorithms are widely used in robot control system.

(xviii) **Virtual Personal Assistant:** ML algorithms are also used in Personal Assistant. It can analyze personal data, process voice requests, automate daily task and can adapt the change in user needs. For example, Alexa by Amazon use all collected data to improve its pattern recognition skills and be able to address new needs on the basis of experience.

(xix) **Traffic Prediction:** This is amongst those application of ML which is used in our daily life. This application is helpful in predicting the traffic data more accurately. It is possible because of ML, which stores all the real-time data and uses it to compute number of cars and their speed on the road and then perform traffic prediction.

(xx) **Self-driving cars:** It is amongst the most common Machine Learning application. ML algorithms are used to train the car models so that it can detect the objects and people while driving.

Description of Machine Learning Techniques for Disease Prediction

The main goal of research study is to collect all the published articles related to this field and conclude about the coverage of research done so far. We focused on the published research from 2017 up to today, review the research done on various machine learning techniques and algorithms used for the efficient prediction of diseases in various healthcare applications.

Min Chen et al.[21] are the first to work on both type of data i.e., structured and unstructured and have proposed a multimodal disease risk prediction model based on CNN. In comparison to other predicting algorithms, this model proposed an accuracy of 94.8% with better convergence speed than CNN-UDRP.

Mehrbakhsh et al.[22]bring techniques based on CART(classification and regression trees), EM(Expectation Minimization), PCA(Principal Component Analysis) and fuzzy rule for diagnosis of disease and obtain good accuracy in prediction. Result indicates that the method of combining fuzzy rule-based along with clustering and PCA are successful in obtaining good accuracy.

Dhiraj et al.[23] suggested a model of disease prediction model based on patients symptoms using the CNN and KNN ML algorithms for accurate disease prediction. Disease symptoms dataset is required for disease prediction. In comparison to KNN algorithm, disease prediction accuracy of CNN is much more (84.5%). Along with the prediction of disease, this proposed system was able to predict the risk (low or high) associated with the disease.

A NB based classifier model was proposed by Venkatesh[24] to predict future health status from heart disease data. Result declares that this method provides an accuracy of approximately 97.12%. The primary aim of this research is to predict a patient’s future health condition.

For predicting fatty liver disease, Mohaimenul Islam[25] developed a ML algorithm based prediction model. Logistic Regression model performance improved with 76.30% accuracy, 74.10% sensitivity and 64.90% specificity in comparison to other classification techniques. Result reveals that it is possible to use Logistic Regression Model as an important tool for clinical decision making.

Pahulpreet and Shriya[26] applied different classification algorithms for early prediction of disease on three different databases and compare the results. From all the comparisons, accuracy of heart disease detection was 87.1% using Logistic Regression, diabetes was 85.71% using SVM and Breast Cancer detection was 98.57% using AdaBoost classifier.

Sayali and Rashmi[27] proposed a method for predicting whether or not a patient has heart disease using NB and KNN algorithm. The high/low risk of illness was also predicted using CNN-UDRP algorithm. Its only drawback is that it uses only structured data whereas Shraddha[28] overcomes the limitation of CNN-UDRP and proposed CNN-UDRP algorithm that uses hospital structured and unstructured data and results that accuracy is more and fast.

Shweta et al.[29] works on different chronic diseases and applied support vector machine, random forest and decision tree to predict whether or not a patient suffers from disease. Chronicle diseases like diabetes, heart disease and liver disease were included and as a result random forest algorithm worked with higher accuracy.

Chronic disease is a major issue of health worldwide. The main cause of death is chronic diseases as per medical standard report. Hence, to reduce the risk of people’s life, Anandajayam[30] works on the analysis of chronical diseases and uses RNN for structured

Received 15 December 2020; Accepted 05 January 2021.

data and CNN for unstructured data. After this, multiple algorithms like SVM, recurrent neural network, K-Nearest Neighbor, naïve bayes and Decision tree were then processed to examine the accuracy rate of disease risk. As a result, performance of RNN algorithm was better in comparison to other algorithms.

A prediction model was suggested by Induja and Raji[31] to reduce the loss of human lives due to cerebral stroke. For training as well as the testing data, ten-fold cross validation was applied and several classification algorithms such as K-nearest neighbor, Decision tree and Naive Bayes were used to predict the risk of stroke. As a result, Decision Tree shows better performance in comparison to all three with an accuracy of 99%.

Ehtisham[32] used SVM and multilinear regression algorithm to predict diseases and perform testing on multiple algorithms such as CNN, decision tree and k-nearest neighbor etc. In comparison, the combination of SVM and multilinear regression provides higher accuracy in the range of 68%-87%.

In combination with APDFS and HLRM, Sandeep Kumar[33] proposed a framework to improve the predictive accuracy of Chronic Kidney Disease by discovering certain characteristics that are important to the diagnosis of Chronic Kidney Disease. As a result, experimental findings showed that as it identifies the disease with 91.6% accuracy, hence the proposed method is efficient.

Theyazn[34] works on an objective to develop the chronic disease surveillance detection system. The dataset was collected from world-wide resources, include ambiguous objects as well. Hence, to remove the ambiguity from dataset, and to boost the performance of a system, the Rough K-means clustering algorithm was used. Different ML algorithms like NB, SVM, KNN and random forest were compared and achieved the best results for diabetic disease classification with Naïve Bayes and RKM (80.55%), while SVM achieved 100% for kidney disease classification in combination with RKM and SVM achieved 97.53% accuracy metric for cancer disease classification along with RKM.

Table 2. Comparison Table of Research Studies on different Chronic Diseases

Reference	Publication /Year	Algorithm Used	Accuracy	Dataset	Disease Predicted	Result
21	IEEE / 2017	CNN based multimodal disease risk prediction	94.8%	Real-life Hospital dataset	cerebral infarction	High or low risk prediction
22	Elsevier/ 2017	CART, EM, PCA with fuzzy rules	-	Public medical data sets taken from	Diabetes and heart disease	Good accuracy in prediction

23	IEEE / 2019	KNN and CNN	84.5%	Downloaded from UCI ML website	General disease	Accuracy of CNN is more in comparison to KNN
24	Springer / 2018	NAaive Bayes technique	97.12%	UCI ML repository	Heart disease	Predicts future health condition of patient
25	EFMI / 2018	RF, SVM, LR, ANN with 10-fold cross validation	76.30%	data were collected from Taipei Medical University Hospital	Liver Disease	A better result is given by the logistic regression technique.
26	IEEE / 2018	Classification algorithms – LR, DT, RF, SVM and adaptive boosting	87.1% in Heart Disease detection 85.71% in Diabetes and 98.57% for Breast Cancer detection	UCI Machine Learning Library	Heart disease, Breast cancer, Diabetes	In order to understand the disease risk prediction, this approach predicts low, high and medium risk of heart disease.
27	IEEE / 2018	CNN-UDRP	65%	UCI Machine Learning Library	Heart disease	Accurate disease risk prediction was achieved.
28	CSEIJ / 2018	CNN-MDRP	94.8%	Hospitals data	Heart disease	Prediction of disease is fast and accurate
29	IEEE / 2018	Random Forest Algorithm	65% 83% 98%	UCI Machine Learning Repository	Liver disease, heart disease and	Random forest algorithm is more accurate

					diabetes	
30	IEEE / 2019	Recurrent neural network and CNN	97.62%	Online dataset from hospitals	cerebral infarction	RNN works better
31	IEEE / 2019	K-nearest neighbor, Decision tree and Naïve Bayes	99%	National Stroke Mortality dataset.	cerebral stroke	Decision tree's performance is much better whereas NB classifiers's performance was poor.
32	IJIRCST/ 2020	support vector machine and multilinear regression algorithm	87%	by collecting patient's symptoms and diagnosis from local hospitals and from open source libraries available online	Heart Disease	Multilinear regression algorithm is better in predicting the chance of heart disease
33	IJPCC / 2020	APDFS and HLRM	91.6%	Medical labs and hospitals	Chronic kidney disease	Model performed well in prior prediction of CKD
34	Hindawi / 2020	RKM with NB, SVM, KNN and RF	80.55% 100% 97.53%	ML repository and Kaggle	Diabetes, Kidney and Cancer	Model successfully diagnosed the chronic diseases

Performance Metrics

Performance metrics [35] is one of the way that researcher use in proposed Machine Learning Algorithm to verify the productivity, performance and efficiency. Some of the measures that validate the work are as follows-

4.1. For Classification Problem

Performance metric which are used to evaluate the classification problem predictions are -

(i) **Confusion Matrix:** Used to determine an accuracy of an algorithm, it is a two dimensional "Actual" and "Predicted" table. By using the sum of the diagonal of a matrix, the total numbers of correctly predicted values are calculated and all the others are considered as incorrect[36]–[38].

Table 3: Confusion Matrix

	Predicted	
Actual	True Positive (TP)	False Negative (FN)
	False Positive (FP)	True Negative (TN)

The terms mentioned in confusion matrix are explained as –

- True Positive: Shows the presence of disease in a patient when it is actually positive.
- True Negative: Shows the absence of disease in a patient when it is actually negative.
- False Positive: Shows the presence of disease in a patient when it is actually negative.
- False Negative: Shows the absence of disease in a patient when it is actually positive.

(ii) **Accuracy**: This is the cumulative number of predictions that are accurate over all other predictions. With the help of a formula, the Accuracy can be calculated as-

$$\text{Accuracy} = \frac{TN+TP}{TN+FP+FN+TP} (1)$$

(iii) **Precision**: It is mainly used in Information Retrieval, it tells us about what proportion of model is actually positive. With the help of a formula, the Precision can be calculated as-

$$\text{Precision} = \frac{TP}{FP+TP} (2)$$

(iv) **Recall / True Positive Rate / Sensitivity:** It tells us about the number of positives returned by the model. With the help of a formula, the Recall can be calculated as-

$$\text{Recall} = \frac{TP}{FN+TP} \quad (3)$$

(v) **Specificity / True Negative Rate:** It tells us about the number of negatives that the model returns. With the help of a formula, the Specificity can be calculated as-

Specificity = $\frac{TN}{FP+TN}$ (4)

(vi) **F1 Score:** It is a single score that represents both recall and precision by giving their harmonic mean. For F1 score, the best value would be 1 and worst would be 0.

F1 Score = 2 * $\frac{Recall\ *Precision}{Precision\ +Recall}$(5)

(vii) **AUC Curve:** AUC means Area under ROC Curve. It is a performance metric used to classify the accuracy of an algorithm.

(viii) **ROC:** ROC is a probability curve whereas AUC calculates the separability that defines a classifier's performance. The plot of a TP rate is drawn against a FP rate in a diagram shown in figure 4. Higher the value of AUC, better will be the model. The quality of ROC Curve is calculated based on the following four constraints[39]–[41].

- If the AUC ranges from 0.9 to 1, then the quality of the test is excellent.
- If the AUC ranges from 0.8 to 0.9, then the quality of the test is Good.
- If the AUC ranges from 0.7 to 0.8, then the quality of the test is Fair.
- If the AUC ranges from 0.6 to 0.7, then the quality of the test is Poor.

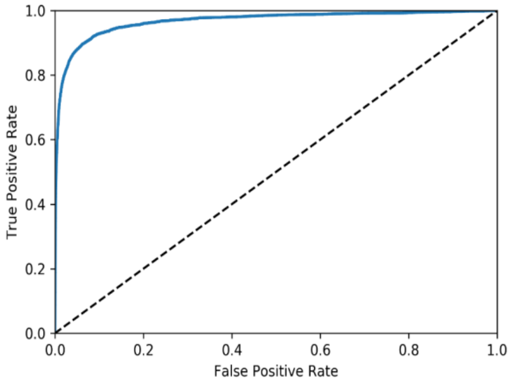

Figure 4. TP rate against FP rate

For Regression Problem

Performance metric that are used to evaluate regression problem predictions are –

(i) **Mean Absolute Error (MAE):** As the name indicates, it is an average of all absolute errors used in regression problems. An average of the absolute difference between the predicted values and actual values is determined. The MAE measurement formula is –

MAE = $\frac{1}{n}\sum|\,Y - \hat{Y}\,|$ (6)

Here, Y = Actual output values

Ŷ= Predicted output values

(ii) **R² or Coefficient of Determination:** It shows the goodness of fit of a set of predicted output values to the actual output values, i.e. it is used to check – how close the data is to the fitted regression line. The R-squared value lies between 0 and 1 where 0 means no-fit and 1 means perfectly-fit. If we denote Yi as the value of the dependent variable observed ,̅Yi as its means and Ŷi as the value fitted then the coefficient of determination is –

$$R2 = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(Yi - \hat{Y}i)^2}{\frac{1}{n}\sum_{i=1}^{n}(Yi - \bar{Y}i)^2} \qquad (7)$$

(iii) **Mean Square Error (MSE):** One of the preferred metrics for regression tasks is MSE or Mean Squared Error. The squared difference between the target value and the predicted value can be calculated as an average.

$$MSE = \frac{1}{n}\sum(Y - \hat{Y})^2 \qquad (8)$$

It penalizes even a small error as it squares the difference that leads to over-estimating how bad the model is.

(iv) **Root Mean Square Error (RMSE)**: RMSE is the most commonly used regression task metric and is the square root of the average square difference between the model's target value and the predicted value. It can be calculated as –

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(Yj - \hat{Y}j)^2} (9)$$

RMSE is extremely affected by outlier values, so before using this metric it is important to remove outliers from the data.

CONCLUSION AND FUTURE SCOPE

As new innovations are expanding to open the new door for decision support system in medical care. These models emphasize on patient's care, accurate diagnosis, correct treatment and helps in medical cost reduction.

In this survey, dataset used by majority of researchers was hospital data that is classified into Structured and Unstructured data. Although, less work is done on both categories of data but various researches from numerous domains are working to achieve disease prediction using both structured and unstructured data.

In this paper, we discusses different ML techniques and their accuracy that other researchers used to diagnose chronic diseases. With the help of such disease prediction model, it is possible to diagnose people suffering from diseases on the basis of their symptoms.

After studying all the mentioned techniques, it has been observed that features and independent variable selection plays an important role in improving the accuracy as well as performance of an algorithm. It has also been observed that single algorithm couldn't

Received 15 December 2020; Accepted 05 January 2021.

performed well in comparison to the combination of multiple algorithms that helps in improving the accuracy. Also, combinations must be used in multiple sequences so that it can be checked that which combination is performing well in comparison to the others in predicting the chronic disease. With the help of a disease prediction system, it is possible to diagnose people based on symptoms, hence selection of correct model is important to perform ideal decision regarding CD diagnosis.

In future, various AI methods like deep learning, cognitive computing can play a vital role in analyzing chronic diseases. Various medical reports like MRI scans, X-rays etc. can be used as a dataset for better accuracy.

REFERENCES

[1] A. M. Minino and S. L. Murphy, “Death in the United States, 2010.,” *NCHS Data Brief*, no. 99, pp. 1–8, 2012.

[2] V. Patel *et al.*, “Chronic diseases and injuries in India,” *Lancet*, vol. 377, no. 9763, pp. 413–428, 2011, doi: 10.1016/S0140-6736(10)61188-9.

[3] S. M. Singh and D. B. Hanchate, “Improving disease prediction by machine learning,” *Int J Res Eng Technol*, vol. 5, no. 6, pp. 1542–1548, 2018.

[4] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, “Disease Prediction by Machine Learning over Big Data from Healthcare Communities,” *IEEE Access*, vol. 5, no. 3, pp. 8869–8879, 2017, doi: 10.1109/ACCESS.2017.2694446.

[5] D. Nagavci, M. Hamiti, and B. Selimi, “Review of prediction of disease trends using big data analytics,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 8, pp. 46–50, 2018, doi: 10.14569/ijacsa.2018.090807.

[6] M. E. Farooqui and D. J. Ahmad, “a Detailed Review on Disease Prediction Models That Uses Machine Learning,” *Int. J. Innov. Res. Comput. Sci. Technol.*, vol. 8, no. 4, pp. 326–330, 2020, doi: 10.21276/ijircst.2020.8.4.14.

[7] M. Chen, Y. Ma, Y. Li, D. Wu, Y. Zhang, and C. H. Youn, “Wearable 2.0: Enabling Human-Cloud Integration in Next Generation Healthcare Systems,” *IEEE Commun. Mag.*, vol. 55, no. 1, pp. 54–61, 2017, doi: 10.1109/MCOM.2017.1600410CM.

[8] P. Shimpi, “A Machine Learning Approach for the,” vol. 6, no. Iccmc, pp. 603–607, 2017.

[9] G. V. Gayathri and S. C. Satapathy, “A Survey on Techniques for Prediction of Asthma,” *Smart Innov. Syst. Technol.*, vol. 159, pp. 751–758, 2020, doi: 10.1007/978-981-13-9282-5_72.

[10] M. Learning, *Machine learning 분야소개및주요방법론학습기본 machine learning 알고리즘에대한이해및응용관련최신연구동향습득*, vol. 45, no. 13. 2017.

[11] C. G. Raji and S. S. Vinod Chandra, “Long-Term Forecasting the Survival in Liver Transplantation Using Multilayer Perceptron Networks,” *IEEE Trans. Syst. Man, Cybern. Syst.*, vol. 47, no. 8, pp. 2318–2329, 2017, doi: 10.1109/TSMC.2017.2661996.

[12] W. Richert and L. P. Coelho, *Building Machine Learning Systems with Python*. .

[13] Ö. Çelik, “A Research on Machine Learning Methods and Its Applications,” *J. Educ. Technol. Online Learn.*, vol. 1, no. 3, pp. 25–40, 2018, doi: 10.31681/jetol.457046.

[14] M. Metcalf, “A first encounter with f90,” *ACM SIGPLAN Fortran Forum*, vol. 11, no. 1, pp. 24–32, 1992, doi: 10.1145/134304.134306.

[15] I. Kecerdasan and P. Ikep, *No 主観的健康感を中心とした在宅高齢者における健康関連指標に関する共分散構造分析Title*. .

[16] B. Nithya and V. Ilango, “Predictive analytics in health care using machine learning tools and techniques,” *Proc. 2017 Int. Conf. Intell. Comput. Control Syst. ICICCS 2017*, vol. 2018-Janua, pp. 492–499, 2017, doi: 10.1109/ICCONS.2017.8250771.

[17] R. Konieczny and R. Idczak, “Mössbauer study of Fe-Re alloys prepared by mechanical alloying,” *Hyperfine Interact.*, vol. 237, no. 1, pp. 1–8, 2016, doi: 10.1007/s10751-016-1232-6.

[18] B. Rao, “Machine Learning Algorithms: A Review,” *Int. J. Comput. Sci. Inf. Technol.*, vol. 7, no. 3, pp. 1174–1179, 2016.

[19] L. P. Kaelbling, M. L. Littman, and A. W. Moore, “Reinforcement learning: A survey,” *J. Artif. Intell. Res.*, vol. 4, pp. 237–285, 1996, doi: 10.1613/jair.301.

[20] G. Winter, “Machine learning in healthcare harvesting of results that a consultant,” vol. 25, no. 2, pp. 100–101, 2019.

[21] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, “Disease Prediction by Machine Learning over Big Data from Healthcare Communities,” *IEEE Access*, vol. 5, no. c, pp. 8869–8879, 2017, doi: 10.1109/ACCESS.2017.2694446.

[22] M. Nilashi, O. bin Ibrahim, H. Ahmadi, and L. Shahmoradi, “An analytical method for diseases prediction using machine learning techniques,” *Comput. Chem. Eng.*, vol. 106, pp. 212–223, 2017, doi: 10.1016/j.compchemeng.2017.06.011.

[23] D. Dahiwade, G. Patle, and E. Meshram, “Designing disease prediction model using machine learning approach,” *Proc. 3rd Int. Conf. Comput. Methodol. Commun. ICCMC 2019*, no. Iccmc, pp. 1211–1215, 2019, doi: 10.1109/ICCMC.2019.8819782.

[24] R. Venkatesh, C. Balasubramanian, and M. Kaliappan, “Development of Big Data Predictive Analytics Model for Disease Prediction using Machine learning Technique,” *J. Med. Syst.*, vol. 43, no. 8, 2019, doi: 10.1007/s10916-019-1398-y.

[25] M. Mohaimenul Islam, C. C. Wu, T. N. Poly, H. C. Yang, and Y. C. Li, “Applications of machine learning in fatty live disease prediction,” *Stud. Health Technol. Inform.*, vol. 247, pp. 166–170, 2018, doi: 10.3233/978-1-61499-852-5-166.

[26] P. S. Kohli and A. L. Regression, “2020 IEEE 5th International Conference on Computing Communication and Automation, ICCCA 2020,” *2020 IEEE 5th Int. Conf. Comput. Commun. Autom. ICCCA 2020*, pp. 1–4, 2020.

[27] S. Ambekar and R. Phalnikar, “Disease Risk Prediction by Using Convolutional Neural Network,” *Proc. - 2018 4th Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2018*, pp. 1–5, 2018, doi: 10.1109/ICCUBEA.2018.8697423.

[28] A. Agrawal, H. Agrawal, S. Mittal, and M. Sharma, “Disease Prediction Using Machine Learning,” *SSRN Electron. J.*, pp. 6752–6757, 2018, doi: 10.2139/ssrn.3167431.

[29] S. Ganiger and K. M. M. Rajashekharaiah, “Chronic Diseases Diagnosis using Machine Learning,” *2018 Int. Conf. Circuits Syst. Digit. Enterp. Technol. ICCSDET 2018*, pp. 1–6, 2018, doi: 10.1109/ICCSDET.2018.8821235.

[30] P. Anandajayam, S. Aravindkumar, P. Arun, and A. Ajith, “Prediction of chronic disease by machine learning,” *2019 IEEE Int. Conf. Syst. Comput. Autom. Networking, ICSCAN 2019*, pp. 1–6, 2019, doi: 10.1109/ICSCAN.2019.8878724.

[31] S. N. Induja and C. G. Raji, “Computational Methods for Predicting Chronic Disease in Healthcare Communities,” *2019 Int. Conf. Data Sci. Commun. IconDSC 2019*, pp. 1–6, 2019, doi: 10.1109/IconDSC.2019.8817044.

[32] M. E. Farooqui and D. J. Ahmad, “Disease Prediction System using Support Vector Machine and Multilinear Regression,” *SSRN Electron. J.*, pp. 331–336, 2020, doi: 10.2139/ssrn.3673232.

[33] S. Hegde and M. R. Mundada, “Early prediction of chronic disease using an efficient machine learning algorithm through adaptive probabilistic divergence based feature selection approach,” *Int. J. Pervasive Comput. Commun.*, 2020, doi: 10.1108/IJPCC-04-2020-0018.

[34] T. H. H. Aldhyani, A. S. Alshebami, and M. Y. Alzahrani, “Soft Clustering for Enhancing the Diagnosis of Chronic Diseases over Machine Learning Algorithms,” *J. Healthc. Eng.*, vol. 2020, 2020, doi: 10.1155/2020/4984967.

[35] N. M. J. Kumari and K. K. V. Krishna, “Prognosis of Diseases Using Machine Learning Algorithms: A Survey,” *Proc. 2018 Int. Conf. Curr. Trends Towar. Converging Technol. ICCTCT 2018*, pp. 1–9, 2018, doi: 10.1109/ICCTCT.2018.8550902.

[36] M. Kaur, H. K. Gianey, D. Singh, and M. Sabharwal, “Multi-objective differential evolution based random forest for e-health applications,” *Mod. Phys. Lett. B*, vol. 33, no. 5, 2019, doi: 10.1142/S0217984919500222.

[37] J. P. Singh and R. S. Bali, “A hybrid backbone based clustering algorithm for vehicular ad-hoc networks,” in *Procedia Computer Science*, 2015, vol. 46, pp. 1005–1013, doi: 10.1016/j.procs.2015.01.011.

[38] N. Mittal, U. Singh, and B. S. Sohi, “A Novel Energy Efficient Stable Clustering Approach for Wireless Sensor Networks,” *Wirel. Pers. Commun.*, vol. 95, no. 3, pp. 2947–2971, 2017, doi: 10.1007/s11277-017-3973-1.

[39] P. Gairola, S. P. Gairola, V. Kumar, K. Singh, and S. K. Dhawan, “Barium ferrite and graphite integrated with polyaniline as effective shield against electromagnetic interference,” *Synth. Met.*, vol. 221, pp. 326–331, 2016, doi: 10.1016/j.synthmet.2016.09.023.

[40] G. Sharma, S. Sharma, and S. Gujral, “A Novel Way of Assessing Software Bug Severity Using Dictionary of Critical Terms,” in *Procedia Computer Science*, 2015, vol. 70, pp. 632–639, doi: 10.1016/j.procs.2015.10.059.

[41] M. K. Gupta *et al.*, “Parametric optimization and process capability analysis for machining of nickel-based superalloy,” *Int. J. Adv. Manuf. Technol.*, vol. 102, no. 9–12, pp. 3995–4009, 2019, doi: 10.1007/s00170-019-03453-3.