# THE CUTTING EDGE OF MACHINE LEARNING IN PREDICTING HIGH - RISK DISEASES IN HEALTHCARE SECTOR

*C. Sandhya*, S. Hemalatha*

**Abstract**

The most novel technique of Machine Learning has transformed the conventional statistical methods into a coveted position. The adaption of digital technologies in the health care sector resulted in a substantial increase in electronic data. Due to the massive output of data, medical practitioners face challenges analysing this data accurately and diagnosing the diseases at an early stage. The application of Machine Learning in this context has proved highly useful and rewarding. Mainly the Supervised Machine Learning algorithms like K-Nearest Neighbor (KNN), Support–Vector Machine (SVM), Decision Trees (DT), Logistic Regression (LR), Naive Bayes (NB) and Random Forest (RF) have proved effective in improving the precision of disease prediction.

**Keywords:** Healthcare, Supervised Machine Learning, Disease Prediction.

## I INTRODUCTION

The application of Machine Learning in the computational field has opened new opportunities in the field of computer science. Machine learning is the most novel technology coming under the vast expanse of Artificial Intelligence that helps computers learn things from the available data and then apply that knowledge for problem-solving without any human effort and thus adapt to situations accordingly. It has a wide range of applications, of which disease prediction using health data is a potential area for illustrating the accuracy of machine learning methods. Inferences are drawn from available data by applying statistical models and algorithms.

Department of Computer Science,
Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu, India
*Corresponding Author

There is enormous information about patient health and hence it is complicated for humans to process it using conventional methods. Machine learning is widely used to detect and predict major diseases that pose a high risk to humans like cardiovascular diseases, cancer, diabetes mellitus and diseases that affect the neurons involving cognitive functions like Alzheimer's and movement disorder like Parkinson's disease. Compared to other dominant technologies like Image Processing, Data mining, Big Data analytics, etc. the role of machine learning is gaining great acceptance especially in the field of medical research. It has a high degree of prediction propensity. The three main popular Machine Learning algorithms that are in vogue namely Supervised Learning, Unsupervised Learning, and Reinforcement Learning. Out of this Supervised Learning is very popular. In this approach, datasets with known labels are induced to predicting models to find unlabeled examples. The supervised learning finds its application in bioinformatics, speech recognition, spam detection, and object recognition for vision. Some important algorithms in this category include Linear Regression, K-Nearest Neighbor, Naive Bayes, Random Forest and Support Vector Machine.

An effort is made to critically analyse the role of machine learning in the field of medicine.

## II LITERATURE REVIEW

Newer trends in technology have made complicated issues to be handled with considerable ease and dexterity. Deep Learning and Machine Learning are modifying the conventional methods in problem-solving. Many studies have highlighted the impact of these technologies on

predicting common high-risk diseases with high precision and accuracy. Following is an example of how this can be exploited in the study of chronic diseases.

**(A) Diabetes Mellitus**

Diabetes Mellitus is a metabolic disorder affecting a large population. It is characterised by the body's inability to metabolise glucose. The risk factors include age, obesity, sedentary habits, high blood pressure, hereditary, dietary imbalances. People with diabetes are vulnerable to a sudden worsening of their symptoms, resulting in heart attack, stroke, nerve damage, and nephrological diseases. Currently, the required information for diagnosing diabetes is collected through various tests in hospitals, and appropriate treatment is provided. Healthcare industries create and maintain large volume databases. Using Machine learning, one can explore these vast datasets and reveal hidden patterns to discover knowledge and predict outcomes accordingly. The present model does not provide an accurate method of classification and prediction. This paper [1], envisages a better prediction model for the classification of diabetes. Various machine learning algorithms were used for classification and out of which Logistic Regression showed the maximum accuracy of 96%.

AdaBoost classifier had an accuracy of 98.8 % when the pipeline was applied. The model showed an improvement in diabetes prediction when the existing dataset was replaced with the new one. Further, this work can be extended to find the probability of non-diabetic people developing diabetes in the future. This work [2] applied ensemble techniques, and machine learning classification algorithms were applied to the dataset to predict diabetes. KNN, LR, Gradient Boosting (GB), Random Forest (RF), DT and SVM are used to predict. The accuracy of these models was compared, and the results showed that RF achieved higher accuracy than other models. The following chart illustrates this.

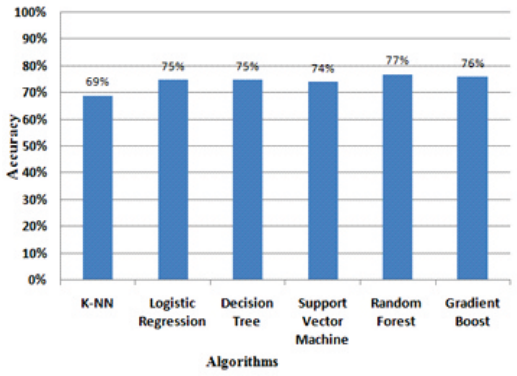


***Fig 1: Accuracy of Machine Learning Methods***

This paper [3] helps detect diabetes and provides a framework to alert medical professionals for timely intervention. In these various models, healthcare and consumer devices were integrated with the cloud to provide medical professionals with accurate and precise readings, enabling them to make informed decisions without delay and least intervention. This study involved the implementation of four main supervised machine learning algorithms. The best performing one was selected and deployed to the framework. The results revealed that SVM- RBF (Support – Vector Machine – Radial Basis Function) showed the highest performance with an accuracy of 83.20 %, sensitivity of 87.20%, and specificity of 79%.

**(B) Parkinson's disease**

Parkinson's disease affects ten million populations worldwide. As identified, the main cause of this disease is the reduced level of the neurotransmitter called dopamine in the area of the brain called substantia nigra. Despite the intensive research, we have not successfully detected the disease's early onset and measures to contain it. At present, the tool available is solely based on the clinical examination of the patients. Various causes have been attributed to the development of this state, including genetics, environmental variants, etc. The currently available methods of imaging like Single – Photon Emission Computed Tomography (SPECT),

Positron Emission Tomography (PET) and Dopamine Active Transfer (DAT)Scan , etc which are highly expensive and requires the expertise of a seasoned specialist in interpreting the results. Early detection of PD is not available. The enormous volumes of patient data are not properly appraised. Genetics involvement in Parkinson's disease is being considered as a useful avenue to tap. This being a novel method with high predication rates, cost-effectiveness, and not involving tedious procedural protocols is of interest for the researchers. Machine Learning has revolutionised the method of detecting and predicting results. Compared to the conventional methods of extricating data, machine learning has proved superior in unfolding hidden patterns.

This study [4] used a Single Layer Neural Network (SLNN), which was used as a classifier and trained by the Wavelet Kernel -Extreme Learning Method (WK-ELM). A genetic algorithm is used for the optimisation of results. The relevant dataset used in this study is derived from the University of California at Irvine. This method yielded the highest classification accuracy of 96.81%. However, this feature can be considered only as a secondary motor symptom that becomes evident in some PD patients later in the disease progression. Results from various experimental models have suggested a key role of alpha-synuclein in the pathogenesis of PD. Besides triggering, it acts as a mediator of disease progression through pathological spreading. Alpha-synuclein (α-Syn) is a major protein involved in Parkinson's disease (PD) pathology. It is abundantly seen in presynaptic vesicles and plays a vital role in neurotransmission. The primary factor concerned with the development of Parkinson's disease is the loss of striatal neuronal cells in the substantia nigra. The misfolding of α-Syn leads to the formation of fibrils that abnormally accumulate and aggregate to form Lewy bodies. Genetic mutations in the snca gene responsible for synthesising α-Syn results in the familial forms of PD and are the basis of sporadic PD. Further exploration to evaluate the involvement of α-Syn in the pathogenesis of PD has triggered several studies in which α -synuclein signatures were extracted from patients with PD and injected in non-human primates [5]. The machine learning approach was employed in this study which established the role of α-Syn in neuro-degeneration. The results showed that in non-human primates, a small amount of α –synuclein has been highly toxic and leads to the death of dopaminergic cells. Proper identification of marker genes has a key role in facilitating the outcome and treatment of the disease.

Though several machine learning algorithms have been proposed in identifying genes that causes disease, only a few are adopted for PD. A novel technique in delineating prospective genes is demonstrated. A feature representation method involving three features (GA, MA, NA) was used to select from protein sequences and then fed to classifier algorithms like SVM, DT, AdaBoost, Xgboost, and Gradient Descent to identify the genes. The most accurate model is selected and used to train the neural network N-semble for accurate prediction. This method was found to be far superior to the other existing methods.

**(C) Heart Diseases**

Heart disease prediction has a pivotal role in dealing with the risk factors inherent in chronic and acute conditions related to the heart. There is a need for better risk prediction models for cardiovascular events since the number of patients increases. Different Machine Learning algorithms have been instituted in studying the disease process and risks involved. Gudadhe [7] et al. proposed a system that uses Support Vector Machine (SVM) and Multilayer Perceptron Neural Network architecture to identify heart disease. Support Vector Machine helps separate each data point by projecting it into higher dimensions, making it superior to other classifier algorithms like Logistic Regression. It does this by adding relevant features to the model. In this study, the database was divided into two categories by using SVM to

detect the presence or absence of heart disease. They achieved an accuracy of 80.41%. The artificial neural network could classify heart disease into five categories with more accuracy of 97.5%. Kanika Pahwa and Ravinder Kumar [8] proposed a hybrid method for selecting the features on the heart disease dataset for prediction. A better feature selection technique called Support Vector Machine – Recursive Feature Elimination (SVM – RFE) was used that eliminates irrelevant and repetitive features. It could reduce the computational time for classification and help in improving the classification accuracy rate. In this study, they applied Naive Bayes and Random Forest on the subset of features for classifying the data set to identify the presence or absence of heart disease.

The results showed much improvement when selected features were applied along with it. Both Naive Bayes and Random Forest achieved a comparable accuracy rate of 84.15% and 84.16%. Marimuthu et al. [9] tried to predict heart diseases using supervised machine learning techniques. In this study, the author used the attributes like age, gender, chest pain type, resting ECG, etc, and subjected them to different Machine Learning algorithms such as Decision Tree (DT), Linear Regression (LR), K-Nearest Neighbors (KNN), and Naive Bayes (NB) classifier algorithms. It was inferred that the LR algorithm proved superior to the other algorithms used with a high accuracy rate of 86.89%. Coronary Artery Disease (CAD) has a high prevalence among major cardiac diseases worldwide. It develops when the coronary arteries become too narrow or cholesterol blockages develop in the walls. CAD can sometimes lead to a heart attack. Coronary Heart Disease is prevalent in the United States. It results in more than 655,000 deaths every year as reported by the Centers for Disease Control and Prevention.

Predictive models built using Machine Learning (ML) algorithms can assist doctors in the timely diagnosis of CAD and can optimise the results. In this study [10] six different machine learning algorithms are used to predict CAD. The Cleveland dataset is used to get patient data. The obtained result can be used as an excellent clinical tool for CAD detection. The six algorithms subjected to scrutiny revealed accuracy above 80 %, while the deep learning neural network algorithm attained the highest accuracy of more than 93%. The performance of the models is as summarised in the following table

| Model | Accuracy | Sensitivity | F1 Score | AUC | Mean |
|---|---|---|---|---|---|
| **Generalized linear model** | 0.8764 | 0.8000 | 0.8786 | 0.883 | 0.85 |
| **Decision Tree** | 0.7978 | 0.7447 | 0.7970 | 0.801 | 0.78 |
| **Random Forest** | 0.8764 | 0.8261 | 0.8751 | 0.880 | 0.86 |
| **Support - Vector Machine** | 0.8652 | 0.7959 | 0.8662 | 0.871 | 0.84 |
| **Neural Network** | **0.9303** | **0.9380** | **0.8984** | **0.796** | **0.88** |
| **k- Nearest neighbor** | 0.8427 | 0.7872 | 0.8419 | 0.847 | 0.83 |

***Table 1: Performance metrics of the ML models applied on the Cleveland Heart Disease Dataset***

## III CONCLUSION

Machine Learning over contemporary algorithms enabled early detection and prediction of major high-risk chronic diseases like heart disease, cancer, diabetes mellitus, kidney diseases, and some neurodegenerative disease like Parkinson's disease. The literature study has identified the usefulness of different classification algorithms like Logistic Regression (LR), Support – Vector Machine (SVM), K-Nearest Neighbor (KNN), Random Forest (RF) and Gradient Boosting (GB) , etc. Even though these algorithms have been successfully applied in different scenarios, Support – Vector Machine and Random Forest has shown considerable performance. The superiority is shown by applying the techniques in combination (hybrid) for creating predictive models is worth mentioning.

## IV FUTURE SCOPE FOR RESEARCH

In the future, there is a need for designing more complex Machine Learning algorithms Which can fine-tune the performance of algorithms used for the early detection and prediction of diseases. Learning models should be calibrated more precisely after the training phase to improve performance. Datasets used should be expanded on different demographics to avoid overfitting and elevate the deployed models' accuracy levels. Finally, the best approach should be chosen that enables appropriate feature selection methods that boost the learning models' performance.

## REFERENCES

[1] Mujumdar, A., & Vaidehi, V. (2019). Diabetes Prediction using Machine Learning Algorithms. Procedia Computer Science, 165, 292-299.

[2] Mitushi Soni , Dr. Sunita Varma, 2020, Diabetes Prediction using Machine Learning Techniques, International Journal of Engineering Research & Technology (IJERT) (Volume 09, Issue 09 (September 2020)

[3] Ramesh, Jayroop & Aburukba, Raafat & Sagahyroon, Assim. (2021). A remote healthcare monitoring framework for diabetes prediction using machine learning. Healthcare Technology Letters.8. 10.1049/htl2.12010.

[4] Derya Avci, Akif Dogantekin, "An Expert Diagnosis System for Parkinson Disease Based on Genetic Algorithm-Wavelet Kernel-Extreme Learning Machine", Parkinson's disease, Article ID 5264743, 9 pages, 2016.

[5] Bourdenx M, Nioche A, Dovero S, Arotcarena ML, Camus S, "Identification of distinct pathological signatures induced by patient-derived α-synuclein structures in non-human primates"- Sci Adv. May 2020

[6] Arora Priya, Malhi Avleen Ashutosh Mishra, "N-semble-based method for identifying Parkinson's disease genes" - Neural Computing & Applications - April2021

[7] M. Gudadhe, K. Wankhade and S. Dongre, "Decision support system for heart disease based on support vector machine and Artificial Neural Network," 2010 International Conference on Computer and Communication Technology (ICCCT), Allahabad, Uttar Pradesh, 2010, pp. 741-745

[8] K. Pahwa and R. Kumar, "Prediction of heart disease using hybrid technique for selecting features," 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON), Mathura, 2017, pp. 500-504

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

[10] Akella, Aravind, and Sudheer Akella. "Machine learning algorithms for predicting coronary artery disease: efforts toward an open source solution." Future science OA vol. 7, 6 FSO698. 29 Mar. 2021, doi: 10.2144/fsoa-2020-02067