A Machine Learning Approach for Predicting Onset and Progression-Towards Early Detection of Chronic Diseases

Vaibhav Kumar[1], Bhagwinder Singh[2], Shilpi Sharma[3], Dolly Sharma[4]

[1-4]Amity School of Engineering and Technology, Amity University Uttar Pradesh
{vaibhavkumar.vk83@gmail.com, bhagwinder.singh25@gmail.com, ssharma22@amity.edu, dsharma17@amity.edu,}.

Abstract

The paper suggests utilizing data mining and machine learning approaches to diagnose chronic diseases due to their cost-effectiveness and ability to analyse large amounts of patient data to identify patterns that may not be detectable by human experts. This enables early detection of diseases, which can improve patient outcomes. Using machine learning techniques can facilitate delivery for customized medical care to individual patients. This approach allows healthcare providers to develop treatment plans that are tailored to the specific needs and circumstances of each patient, resulting in improved outcomes.In modern times, people are exposed to environmental conditions and adopt lifestyles that make them susceptible to various chronic illnesses. Early detection of illness is crucial, but accurate prediction based solely on symptoms is challenging for doctors. Data mining can aid in disease prediction by finding hidden patterns in the vast amount of medical data generated each year. With the increasing amount of medical data available, accurate analysis can greatly benefit patient care. We propose a healthcare model that uses mining algorithms such as K nearest neighbours, Decision Trees and Logistic Regression, along with a dataset of disease symptoms. Each model also takes into account a person's lifestyle habits and medical check-up information for accurate disease prediction. Logistic Regression has a prediction accuracy of 95.6% for Heart disease, which is higher than the accuracy of Naive Bayes algorithm. On the other hand, the Decision Tree algorithm has an accuracy of 99% for predicting patients with thyroid disease. It is noteworthy that among patients who do not have any thyroid problems, Logistic Regression has an accuracy of 81.16% in predicting those with hyperthyroidism or hypothyroidism.

Keywords :- Chronic Disease, Machine Learning, Logistic Regression, Decision Tree
DOINumber:10.48047/nq.2022.20.22.NQ10187 **NeuroQuantology2022;20(22):2000-2007**

Introduction

Artificial Intelligence has enhanced the computer's intelligence and capacity to think. Within AI research, machine learning is considered a subfield that many experts believe is essential for creating intelligence. There are several MLmethod available, including Unsupervised, supervised, evolutionary learning, and deep learning. These techniques enable the fast and efficient classification of lots of data. To accurately predict chronic diseases, mining algorithms

such as k-nearest neighbour, logistic regression, and decision tree are commonly used for big data classification. Due to the continuous increase in medical data, predicting the correct disease is a crucial task, and the process of handling and mining big data is extremely important. Consequently, classification of large datasets using machine learning has become an easy and efficient solution.

Currently, diabetes is a widely known term and presents a significant challenge for both third world countries [1]. The pancreas produces the Insulin, which allows glucose to enter the bloodstream from food. Malfunctioning of the pancreas and subsequent lack of insulin results in diabetes, which can lead to serious health problems such as coma, renal and retinal failure, cardiovascular dysfunction, joint failure, and more [2]. Studies have shown that diabetes is on the rise, with the percentage of adults over 18 years old living with diabetes increasing from 4.7% to 8.5% between 1980 and 2014 [3]. As of 2017, there were more than 450 million humans are living in diabetes worldwide, and this figure is to increase by 695 million by 2047.Furthermore, research indicates that the severity of diabetes is also increasing, with 5 million people worldwide affected, and the number expected to riseby 25% in 2030 and 50% by 2045 [4], respectively. While there is currently no permanentremedy for diabetes, early prediction and management can help control and prevent its development.

Identifying heart disease can be challenging due to numerous risk factors that contribute to its development, consistingsugar, increase inBP, increase in cholesterol, irregular pulse, and more. To know the extent of heart disease in humans, data mining techniques have been utilized. The severity of the disease is categorized using various methods, such as k nearest neighbours Algorithm , decision trees, and naive bayes [5].

LITERATURE REVIEW

This paper highlights that diabetes is a serious global health issue that can cause various health complications, including blindness. To accurately diagnose the disease, the authors have utilized machine learning techniques, which are effective and flexible in predicting diabetes in patients. Their primary aim is to produce a mechanism that will precisely detect presence for diabetes in individuals. The study utilized four different algorithms - decision tree, naïve bayes, and svmand matched their performance, which was 85.1%, 77.3%, and 77.3% respectively. Additionally, the authors employed the ANN algorithm to evaluate the performance of the network in correctly classifying diabetes. And alsomatched recall, precision, , f1 score, and accuracy.[6].

Objective of thisanalysis is to address the crucial importance of the heart in living organisms and the need for accurate predictions and diagnoses of heart-related diseases, as they can lead to death. To this end, the authors use machine learning and artificial intelligence to predict heart diseases. They evaluate the accuracy of these techniques using knearest neighbours, decision trees, Linear Regression, and Support vector machine algorithms. The comparison in the performance of the algorithms and find that k-nearest neighbor has an accuracy of 87%, followed by SVM with 83%, decision tree with 79%, and linear regression with 78%. [7]

This paper explores the use of different classification algorithms, including decision tree, svm, artificial neural network, and k nearest neighbour algorithm, to predict thyroid disease. The authors obtained a dataset from UCI Repository and performed classification and prediction using these algorithms. They evaluated the accuracy of each algorithm and compared them to determine the best technique with high accuracy.[8]

The paper focuses on diagnosing thyroid nodules as either mild or malignant by

examiningthe pictures of ultrasound, utilizing both radio and Deep Learning methods. The research compares the two approaches and concludes that deep learning performs better. The radiomics method achieves a prediction accuracy of 66.81%, and sensitivity of 51.19% also specificity of 75.77%. Meanwhile, the deep learning method attains evaluation indices of 74.69%, 63.10%, and 80.20%. The findings announce that DL approach is superior than radiomics method.[9]

The paper presents a Heart Disease Prediction System (HDPS) and compares it to KNN, SVM, Random Classifier, and Decision Trees [10]. The outcome shows the decision tree provides most accurate prediction, with an improvement accuracy of 98.85% compared to the other methods. Therefore, the Decision Tree Machine Learning method is recommended for the HDPS.

The authors of the study examined, compared some classification models likenaive bayes, decision trees, neural networks, and radial basis network. They found that all models achieved significant accuracy, but the tree model performed best. The research used 29 attributes in the dataset and applied feature selection techniques such as ChiSquare.Additionally, unsupervised coated filters were used to convert continuous values into nominal values, resulting in a reduction of the 29 attributes to 10 attributes[11].

The mechanismof the diagnosinglevels of heart disease with k star algorithm. The research explains how it used the Learning Vector Quantification Neural System Calculation to output the likelihood of heart disease based on 13 clinical data points[12]. The neural system they developed can accurately showif a person has coronary heart disorder or not, and it has been evaluated using various performance measures.

In reference [13], a systematic approach was proposed to diagnose thyroid disease earlier using a neural network and the backpropagation algorithm. The neural network uses the backpropagation of error to predict the likelihood of thyroid disease. The neural network is trained with empirical data and tested usedobservations not consideredin training process. The output indicates the neural network works in coordination with the data and can be used as a substitute for prior disease predictions, indicating an advanced approach to disease diagnosis.

Sustainable development has become a widely adopted concept that aims to address the increasing environmental and socio-economic concerns[14]. It requires changes in the behavior of individuals, institutions, and businesses. By adopting sustainable development practices, we can better understand our impact on the environment by assessing the overall consequences of our actions.

The COVID-19 pandemic, which is transmitted through contact with infected individuals, has caused significant economic disruption and human suffering worldwide. It has created a high level of uncertainty globally[15]. The ongoing crude oil price war between the world's two largest exporters, Saudi Arabia and Russia, has exacerbated the situation further. This uncertainty has led to a pause in major investment decisions across various industries. The impact of the pandemic highlights the importance of prioritizing human health over financial concerns.

METHODOLOGY

A. Existing system

There are currently various analyses available that focus on specific diseases. For example, if a someone wants its diabetesto be analyzed, and if preference is to analyze stroke disorder, it should use a different one. It can be time-consuming process. Moreover, if a user has multiple diseases, but the existing system can only predict one disease, it could lead to a higher mortality rate as the other disease may go undetected.

B. Proposed system

The development of a multi-disease prediction model has made it possible to predict more than one disease simultaneously, eliminating the need for users to navigate through multiple models for different diseases. This approach saves time, and as it predicts multiple diseases at once, it has the potential to reduce the mortality rate, as depicted in Figure 1.

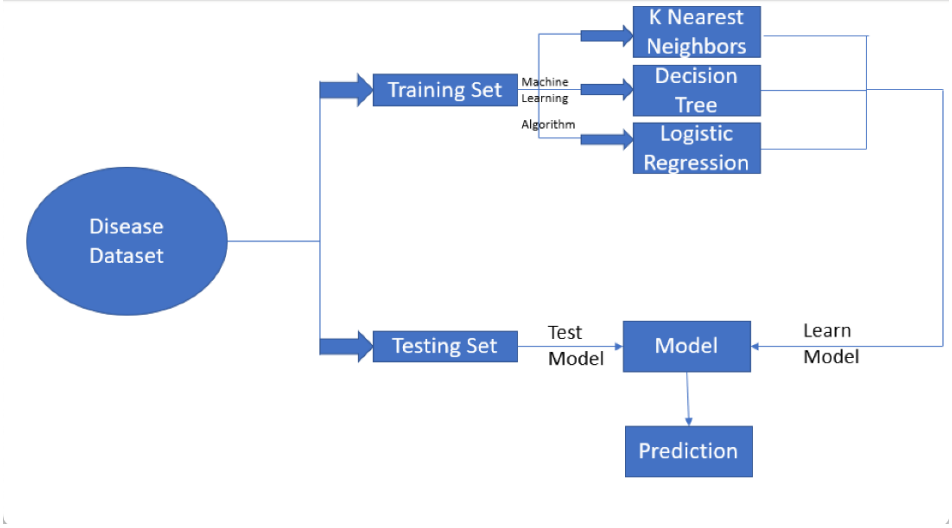

Figure1:A common sequence of developing a machine learning model

C. Data collection

To identify the disease, data collection was conducted from the internet using real symptoms of the disease, and no fake values were included. The dataset were obtained from kaggle website.

D. Data preprocessing.

Prior to inputting the data into the prediction model, certain steps are taken to clean and preprocess the data.

- To ensure completeness of data and enable analysis without gaps or missing information, the process of identifying missing values in a dataset and filling them with the mean or mode of the available data is undertaken
- Standardizing the data .
- Splitting the dataset into training and testing sets.
- We dropped some feature in case of thyroid dataset

E. Building Model.

We introduce a novel prediction model that utilizes different machine learning algorithm and multiple data source to forecast the risk of limited number of chronic diseases. This model has demonstrated improved accuracy compared to existing methods.

The project utilizes classification techniques as a part of data mining process, specifically for categorizing categorical data. The classification process involves two main stages, namely training and testing. In the training stage, pre-determined data and their corresponding class labels are used for classification, which is known as supervised learning. The diagram illustrates the preparation and testing phases of the classification process. The training phase involves using training tuples, while the testing phase involves using test data tuples to calculate the accuracy of the classification rule. Assuming the classification rule's accuracy on the testing data is satisfactory, it can be used for classifying unmined data.

For diabetes dataset we have used different algorithm to get higher accuracy Logistic

Regression With l2 regularization outperform Random forestby a very small margin.

In figure 2 the accuracy of different model on diabetes dataset.

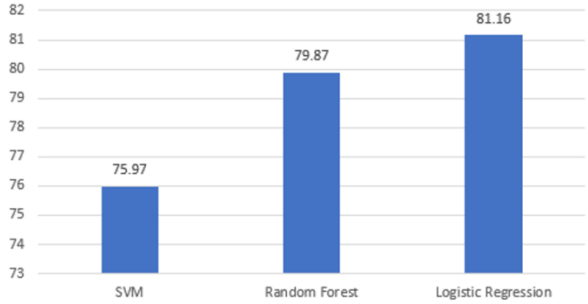

Figure 2: Percentage accuracy of algorithms on Diabetes dataset

For the thyroid dataset, we removed certain features, such as measured_tsh, measured_t3, measured_TT4, measured_T4U, measured_FTI, and measured_TBG, as they were contributing redundant information to the data.

We applied both 'gini index' and ‘entropy’ as impurity measures and found that the 'entropy' measure was more effective in assessing the degree of uncertainty in the dataset, leading to better accuracy in predicting outcomes.

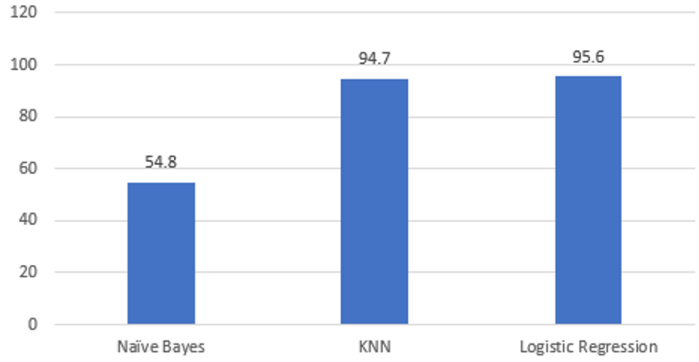

Figure 3: Performance of different ML techniques on predicting Stroke

We employed various versions of knearest neighbours, naive bayes, and random forest algorithms to create a model for predicting strokes. Although Gaussian Naive Bayes failed to establish a precise correlation between the independent and dependent variables, Logistic Regression and Decision Tree models yielded comparable results shown in figure 3.

RESULT

We have successfully applied data mining methods to attain high levels of accuracy, precision, and recall. While these measures are commonly used in information retrieval, they are also pertinent to other measures like sensitivity and specificity, which is why we are taking them into account in this context.

Figure 4,The accuracy of all the three models are shown pictorially. These models can easily be used by patients as well as medical practitioners to be sure before treating a patient. Healthcare professionals can use this as a second opinion for confirmation of a disease.

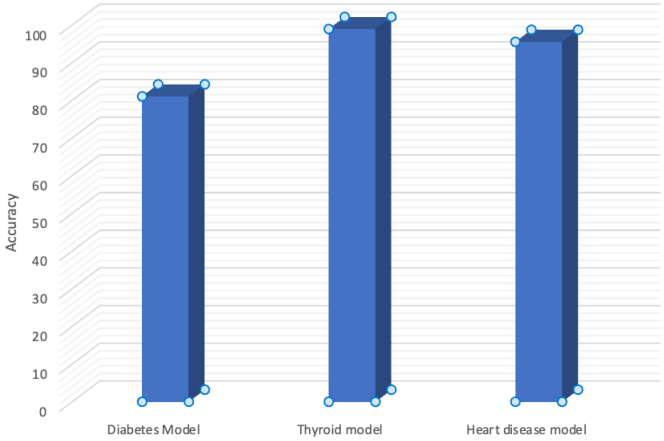

Figure 4:Bar graph depicting the accuracy of disease prediction models

Plots in figure 5 clearly shows that individuals with T3 levels exceeding 6 are highly likely to have Hyperthyroidism, while those with Tsh levels over 180 can be easily identified as Hypothyroid patients a simple if-else condition could be used to determine the presence of these diseases.

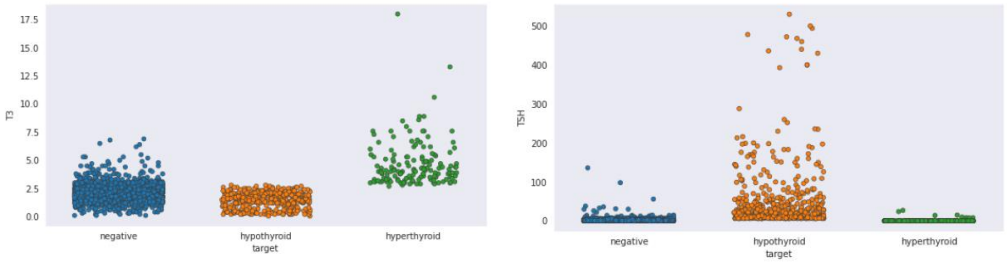

Figure 5: Plots depicting dependency of target variable on T3 and Tsh levels

Once we had cleaned the data, we applied various data mining techniques and found that the Decision Tree Classifier with the 'entropy' criterion outperformed all other algorithmsas shown in figure 6.It is clear from the bar chart given underneath that the algorithms namely svm, logistic regression and decision tree performed fairly well on this dataset but Decision Tree outperforms by a very small margin.

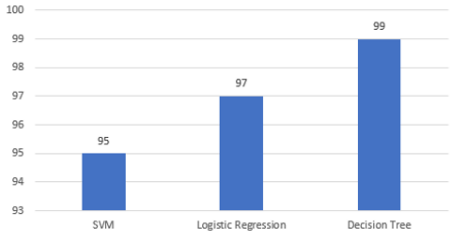

Figure 6: Accuracy of prediction for different machine learning algorithms

Figure 7: Representation of the correlation between variables that contribute to aStroke

Figure 7 verifies a strong association between a person's age and the likelihood of experiencing a stroke, and age appears to be linked to both hypertension and the patient's BMI levels. It signifies that these attributes are very important in prediction of the disease.

CONCLUSION

The primary goal of this system is to predict diseases based on a user's symptoms. The user provides their symptoms as input, and the system generates a prediction of the disease as output. The system was implemented u, and it has an average prediction accuracy probability of 100%. It is designed to be user-friendly and easy to use.

The aim of the review is assess its effectiveness and cons of software used in medical industry, with the aim of informing future developers of disease predictability software and promoting personalized patient care. The software in question predicts patient diseases by analyzing user symptoms. This review aims to provide insights that will facilitate the development of better disease prediction software in the future.

REFERENCES

[1] A. Misra, H. Gopalan, R. Jayawardena, A. P. Hills, M. Soares, A. A. RezaAlbarrán, and K. L. Ramaiya, “Diabetes in developing countries,” Journal of Diabetes, vol. 11, no. 7, pp. 522-539, Mar. 2019

[2] R. Vaishali, R. Sasikala, S. Ramasubbareddy, S. Remya, and S. Nalluri, “Genetic algorithm based feature selection and MOE Fuzzy classification algorithm on Pima Indians Diabetes dataset,” in Proc. International Conference on Computing Networking and Informatics, Oct. 2017, pp. 1-5.

[3] Emerging Risk Factors Collaboration and other, “Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies,” The Lancet, vol. 375, no. 9733, pp. 2215-2222, Jul. 2010.

[4] N. H. Choac, J. E. Shaw, S. Karuranga, Y. Huang, J. D. R. Fernandes, A. W. Ohlrogge, and B. Malandaa, “IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045,” Diabetes Research and Clinical Practice, vol. 138, pp. 271-281, Apr. 2018.

[5] M. Durairaj and V. Revathi, ``Prediction of heart disease using back propagation MLP algorithm, ''back propagation MLP algorithm,'' Int. J. Sci. Technol. Res., vol. 4, no. 8, pp. 235_239,2015.

[6] A. Jabeen, N. Ahmad, K. Raza, Machine learning-based state-ofthe-art methods for theclassification of RNA-Seq data, in: N. Dey, A. Ashour, S. Borra (Eds.), Classification inBioApps. Lecture Notes in Computational Vision and Biomechanics, vol. 26, Springer,Cham, 2018.

[7] Intelligence System for Diagnosis Level of Coronary Heart Disease with K-Star Algorithm WihartoWiharto, MCom1,2, Hari Kusnanto, DrPH3 , HeriantoHerianto, DrEng2.

[8] Ramalingam, V. V., Dandapath, A., & Raja, M. K. (2018). Heart disease prediction using machine learning techniques: a survey. International Journal of Engineering & Technology, 7(2.8), 684- 687.

[9] Priyanka Sonar, Prof. K. JayaMalini," DIABETESPREDICTION USING DIFFERENT MACHINE LEARNINGAPPROACHES", 2019 IEEE ,3rd International Conference on Computing Methodologies and Communication (ICCMC).

[10] Archana Singh ,Rakesh Kumar, "Heart Disease Prediction Using Machine Learning Algorithms", 2020 IEEE, International Conference on Electrical and Electronics Engineering (ICE3).

[11] Ankita Tyagi and Ritika Mehra. (2018)."Interactive Thyroid Disease Prediction System using Machine Learning Techniques" published on ResearchGate.

[12]YongFeng Wang,(2020). "Comparison Study of Radiomics and Deep-Learning Based Methods for Thyroid Nodules Classification using Ultrasound Images" published on IEEEAccess.

[13]S. Sathya Priya, Dr. D. Anitha" Survey on Thyroid Diagnosis using Data Mining Techniques" International Journal of Advanced Research in Computer and Communication Engineering Vol. 6, Special Issue 1, January 2017.

[14]Gulati, Saksham, and Shilpi Sharma. "Challenges and responses towards sustainable future through machine learning and deep learning." Data Visualization and Knowledge Engineering: Spotting Data Points with Artificial Intelligence (2020): 151-169.

[15]Tanmay, Tushar, Akanksha Bhardwaj, and Shilpi Sharma. "The Economic and Technical Factors During Coronavirus Pandemic in affected countries." 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence). IEEE, 2021.