Prediction of Multiple Diseases Using Machine Learning

Sakshi Kumari[1] and Shubham Kumar Singh[2]

[1]Sathyabama Institute of Science and Technology, Jeppiaar Nagar, SH 49A, 600119, Chennai, Tamil Nadu

Article Information

Received: 01-01-2022
Revised: 01-01-2022
Published: 01-01-2022

Keywords:
Machine learning, *Heart disease*, *Parkinson Disease*, *Breast Cancer*, *Lung Cancer*, *Diabetes*

Correspondence Email:
shubhamkumarsingh1017@gmail.com

sakshisingh748894@gmail.com

Abstract

Due to machine learning progress in healthcare communities, accurate study of medical data benefits patient care, early disease recognition and community services. Machine learning and Artificial Intelligence are playing a big role in today's world. From self-driving cars to medical fields, we can find . machine learning and Artificial Intelligence role everywhere. The medical industry generates a huge amount of patient data which can be processed in a lot of ways. So, with the help of machine learning, we have created a Prediction System that can detect more than one disease at a time. Many of the existing systems can predict only one disease at a time and that too with lower accuracy. Lower accuracy can seriously put a patient's health in danger. We have considered five diseases for now and in the future, many more diseases can be added. The user has to enter various parameters of the disease and the system would display the output whether he/she has the disease or not. This project can help a lot of people as one can monitor the persons' condition and take the necessary precautions thus increasing the life expectancy.

1. Introduction

In this digital world, data is an asset, and huge data was generated in every fields. In the healthcare domain, data encompasses a comprehensive set of information specifically related to patients. level of insights can be drawn from this data. So, by using advanced machine learning techniques and this data we have decided to come up with project 'Prediction of Multiple Disease using Machine learning'. Our project is combination of five machine learning models that are going to identify patients having heart disease, Parkinson Disease, Breast Cancer, Lung Cancer and diabetes at early stage, so that the patients having risk to particular disease will get treatment first. For our project we firstly understood problem statement and determined what type of data will be required for our project. We collected five different datasets for our five machine learning models from Kaggle. After collecting data, we analyzed data properly and visualized it for better understanding. Then we cleaned the data by imputing null values, encoding categorical features. Next step was we split dataset intro training and testing set such that we used 80% data for training machine learning model and remaining 20% data is for testing our machine learning model. After that we used several classification algorithms like Logistic regression, Random Forest classifier, Decision Tree, Support Vector Machine classifier, Naïve Bayes classifier on all five datasets. Out of these algorithms we found that Random Forest classification algorithm was

performing better than others for all five datasets. Using Random Forest algorithm we got 98.51% testing accuracy on Lung Cancer datasets ,97.50% accuracy on heart disease dataset ,98.72% testing accuracy on Parkinson Disease dataset and 98.71% accuracy on Breast Cancer disease dataset and 80.55% testing accuracy on diabetes dataset. To implement multiple disease prediction systems, we are going to use machine learning algorithms, and Django. Python pickling is used to save the behaviour of the model. Analyzing a system is crucial because it helps us look at all the factors causing a disease. By considering every parameter, we can detect diseases more efficiently and accurately. Once we've figured out the best way the system works, we save its behavior in a Python pickle file.

1.1 Literature Review

[1] According to the paper focuses about as diabetes is one of the dangerous diseases in the world, it can cause many varieties of disorders which includes blindness etc. In this paper they have used machine learning techniques to find out diabetes disease as it is easy and flexible to forecast whether the patient has illness or not. Their aim of this analysis was to invent a system that can help the patient to detect the diabetes disease of the patient with accurate results. Here they used mainly five main algorithms Decision Tree, Naïve Bayes, logistic regression, Random forest and SVM algorithms and compared their accuracy respectively.

[2] The main aim of the paper is, as heart plays an important role in living organisms. So, the diagnosis and prediction of heart related disease should be perfect and correct because it is very crucial which can cause death cases related to heart. So, Machine learning and Artificial Intelligence supports in predicting any kind of natural events. So, in this paper they calculate accuracy of machine learning for predicting heart disease using k-nearest neighbour, decision tree, linear regression and SVM by using UCI repositor dataset for training and testing. They also compared the algorithm and their accuracy SVM 83 %, Decision tree 79%, Logistic regression 78%, random forest 87%.

[3] Parkinson Disease is a brain neurological disorder. It leads to shaking of the hands, body and provides stiffness to the body. No proper cure or treatment is available yet at the advanced stage of Parkinson Disease. Treatment is possible only when done at the early or onset of the disease. These will not only reduce the cost of the disease but will also possibly save a life. Most methods available can detect Parkinson in an advanced stage; which means loss of approx. 60% dopamine in basal ganglia and is responsible for controlling the movement of the body with a small amount of dopamine. Over 145,000 individuals in the U.K. are found to be suffering alone, while in India, nearly one million people are affected by this disease. Its rapid spread is a growing concern globally.

[4] This research paper explores using machine learning to predict diabetes in a healthy population. The goal is to identify individuals at high risk so that timely interventions can prevent future complications. The researchers used data from the San Antonio Heart Study and developed a model to predict the long-term development of type-2 diabetes. They faced a common machine learning challenge of dealing with an unbalanced dataset, where the model could be biased toward the majority class. To address this, they balanced the classes to create unbiased models. By adjusting the classification threshold, they optimized the model to better identify true positive cases. The study achieved a validation accuracy of 84.1% with a recall rate of 81.1% over multiple iterations. The findings suggest that high glucose levels observed during a specific test can strongly indicate the potential risk of developing type 2 diabetes in the future.

[5] Applying Machine learning methods in Diagnosing Heart disease for Diabetic Patients This research article sheds light on a model that has suggested a method to demonstrate that mining may assist to retrieve relevant correlation even from features that are not direct indicators of the class they are attempting to predict. In their study, they attempted to predict the probability of developing a heart disease using attributes derived from diabetes diagnosis, and they demonstrated that it is feasible to identify heart disease susceptibility in diabetic patients with fair accuracy. This type of classifier can aid in the early detection of a diabetic patient's sensitivity

[6] to heart disease. Patients is probably cautioned to adjust their life-style due to this. This will prevent diabetes people from developing heart disease, resulting in lower death rates and lower healthcare costs for the state

[5] Matta, Ahmad, Bhattacharya, and Kumar (2022) have worked on using machine learning to improve systems that detect and prevent attacks by botnets. In their study, they analyze network traffic to identify and eliminate botnets effectively. They propose a comprehensive strategy that employs advanced tools for quickly finding and stopping botnet activities. You can find more detailed information about their work in the chapter titled "Advanced Attack Detection and Prevention Systems by Utilizing Botnet" on pages 27–53 of the CRC Press book Real-Life Applications of the Internet of Things: Challenges, Applications, and Progress.

2. METHODOLOGY

A. Machine Learning Models

In the initial phase, the project focuses on defining the problem and selecting suitable machine learning models. Open data sources like Kaggle and the UCI Machine Learning Repository are utilized for data collection. Ensuring both quality and quantity of data is paramount for model accuracy. The collected data undergoes a comprehensive preprocessing step to format it correctly. Analysis involves handling duplicates, missing values, and outliers. Visualization aids in understanding variable relationships, drawing insights, and addressing skewness. The dataset is split into training and testing sets, with 80% allocated for training and 20% for testing. Multiple machine learning algorithms are explored.

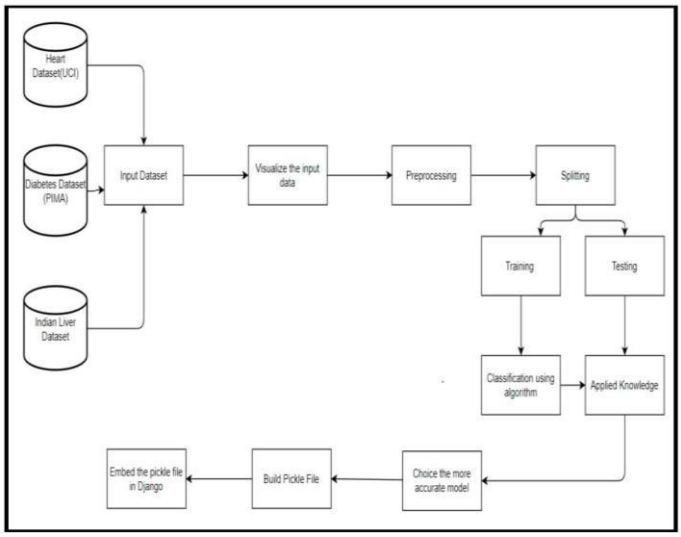

Fig 1. Skema

B. Deploying ML Models in Flask

Following the creation of machine learning models, deployment is executed within the Flask framework of the Python language. The Pickle module is utilized for this purpose, facilitating the serialization and deserialization of machine learning models. The deployment includes a graphical user interface (GUI) for user interaction.

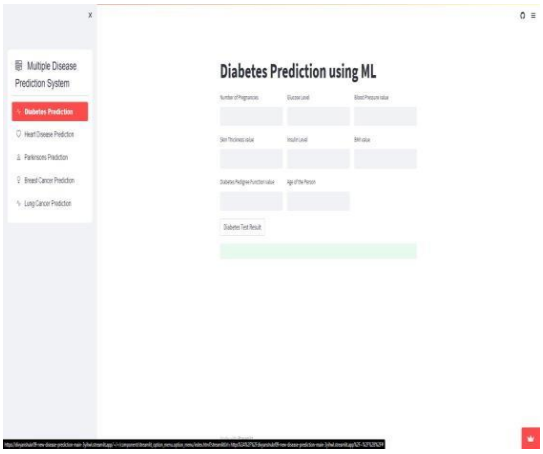

Fig 2. Diabetes User Interface

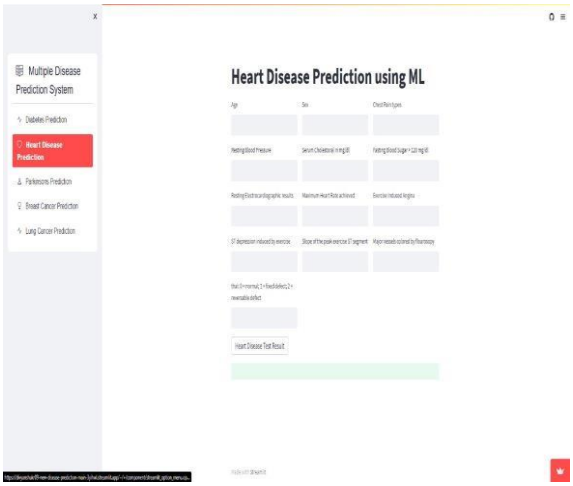

Fig 3. Heart Disease User Interface

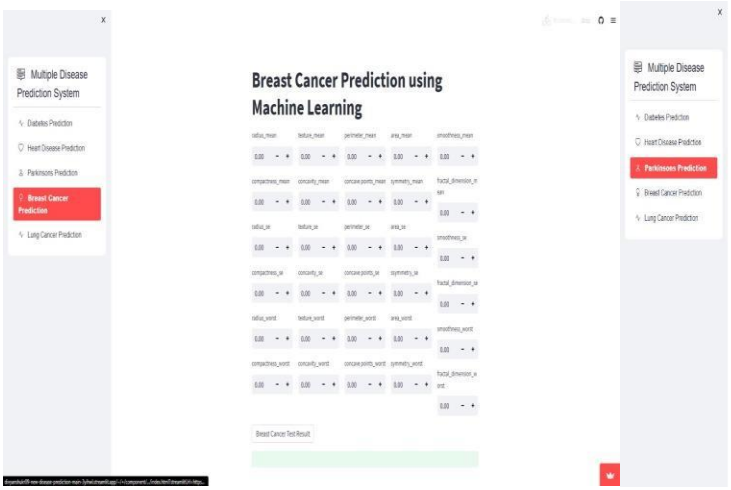

Fig 5.Breast Cancer User Interface

Fig 4 Parkinson's Disease User Interface

Fig 6 Lung Cancer User Interface

Random Forest:

The Random Forest algorithm operates in two phases: creating the forest by combining N decision trees and making predictions for each tree. The process involves selecting random K data points, building decision trees associated with the selected points, determining the number N of decision trees, and repeating the process. Predictions are obtained for each decision tree, and the new data points are assigned to the category with the majority votes.

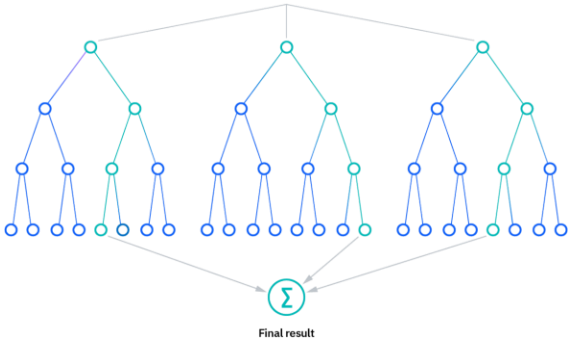

Fig 7. The Random Forest algorithm

Logistic Regression:

Logistic Regression is a supervised machine learning algorithm used for classification, specifically for predicting values which can be categorized (True/False or 1/0). Instead of fitting a regression line, it employs an "S"-shaped logistic function, providing probabilities between 0 and 1. The logistic regression equation is based on the logarithm of the odds of the predicted outcome.

Support Vector Classifier:

Support Vector Machine (SVM) is a supervised machine learning algorithm suitable for classification and regression. SVM is to find a hyperplane in N-dimensional space for distinct classification of data points, with the hyperplane's dimension dependent on the number of features.

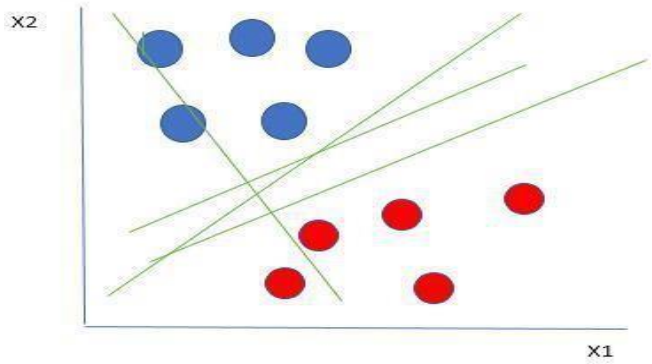

Fig 8. Support Vector Machine (SVM)

Decision Trees:

Decision Trees are a non-linear supervised machine learning algorithm used for both regression and classification tasks. The algorithm makes decisions based on asking a series of questions to classify or predict the target variable. Trees are built by recursively splitting the data based on the most significant feature at each node.

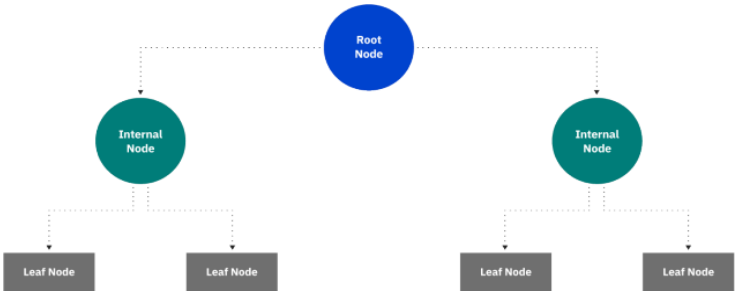

Fig 9. Decision Trees

Naive Bayes:

Naive Bayes is a probabilistic classification algorithm based on Bayes' theorem, assuming independence between features. It's particularly useful for text classification and simple yet effective for various classification tasks. The algorithm calculates the probability of each class given a set of features and selects the class with the highest probability.

3. Conclusions

1. In conclusion, the field of Multiple Disease Prediction using Machine Learning holds tremendous potential for revolutionizing healthcare by enabling early detection, personalized interventions, and efficient resource allocation. Through the integration of datasets and the application of various machine learning algorithms, the progress has been made in predicting diseases such as heart disease, diabetes, lung cancer, Parkinson's disease, and breast cancer.
2. The findings from this project highlight the effectiveness of machine learning algorithms, including logistic regression, decision trees, support vector machines, random forests and naive Bayes in predicting multiple diseases. These algorithms offer unique strengths and advantages, ranging from interpretability and simplicity to handling complex relationships and high-dimensional data. By leveraging features such as clinical records, biomarkers, lifestyle factors, and emerging data sources, more accurate assessments can be achieved.
3. The successful implementation of diseases based on machine learning prediction models requires careful consideration quality of data, model selection, feature engineering, validation, and ongoing monitoring to ensure their reliability and generalizability. Additionally, future research and development are crucial to refine existing algorithms, explore new techniques, validate predictive models in diverse populations, and overcome challenges related to data privacy and ethical considerations.
4. By enabling early intervention, personalized medicine, multiple disease prediction using machine learning has the potential to significantly impact healthcare outcomes. This includes improving patient care, reducing healthcare costs, and enhancing population health management. Furthermore, the integration of machine learning models into clinical decision support systems can empower healthcare providers with actionable insights for informed decision-making.

FUTURE SCOPE

1. Integration of Multi-omics Data: Future research may focus on integrating multiple types of omics data, such as genomics, proteomics, metabolomics, and epigenomics, to enhance the predictive power of machine learning models. This comprehensive approach can provide a more holistic understanding of disease development and progression.
2. Incorporation of Environmental Factors: Expanding machine learning models to include environmental data, such as air quality, pollution levels, lifestyle factors, and socioeconomic indicators can improve disease prediction accuracy. This integration can help identify and assess the impact of environmental risk factors on disease outcomes.
3. Real-Time Data Analysis: Advancements in data processing speed and availability of real-time health data from wearable devices and sensors offer opportunities for real-time disease prediction and

monitoring. Machine learning models can be developed to analyze streaming data and provide timely alerts or interventions for early detection and prevention.

4. Interpretable and Explainable Models: Enhancing the interpretability of machine learning models is crucial for their acceptance and adoption in clinical practice. Future efforts may focus on developing models that provide transparent explanations for predictions, allowing healthcare professionals to understand and trust the decision-making process.

4. References

[1] R. Manne, S.C. Kantheti, Application of artificial intelligence in healthcare: chances and challenges, Curr. J. Appl. Sci. Technol. 40 (6) (2021) 78–89, https:// doi.org/10.9734/cjast/2021/v40i631320.

[2] M. Sivakami, P. Prabhu. Classification of algorithms supported factual knowledge recovery from cardiac data set, Int. J. Curr. Res. Rev. 13(6) 161- 166. ISSN: 2231-2196 (Print) ISSN: 0975-5241 (Online).

[3] M. Sivakami, P. Prabhu. A Comparative Review of Recent Data Mining Techniques for Prediction of Cardiovascular Disease from Electronic Health Records. In: Hemanth D., Shakya S., Baig Z. (eds) Intelligent Data Communication Technologies and Internet of Things. ICICI 2019. Lecture Notes on Data Engineering and Communications Technologies, vol 38. Springer, Cham 477-484. ISSN 2367-4512 ISSN 2367-4520 (electronic), ISBN 978-3-030-34079-7 ISBN 978-3-030-34080-3 (eBook) 2020.

[4] P. Prabhu, S. Selvabharathi. Deep Belief Neural Network Model for Prediction of Diabetes Mellitus. In 2019 3rd International Conference on Imaging, Signal Processing and Communication, ICISPC 2019 (pp. 138–142) Institute of Electrical and Electronics Engineers Inc. ISBN:9781728136639. 2019.

[5] N. Jothi, N.A. Rashid, W. Husain, Data mining in healthcare – A review, Procedia Comput. Sci. 72 (2015) 306–313.

[6] H. Polat, H. Danaei Mehr, A. Cetin. Diagnosis of chronic kidney disease based on support vector machine by feature selection methods, J. Med. Syst. 41(4) 2017 55.

[7] K.B. Wagholikar, V. Sundararajan, A.W. Deshpande, Modeling paradigms for medical diagnostic decision support: a survey and future directions, J. Med. Syst. 36 (5) (2012) 3029–3049.

[8] E. Gürbüz, E. Kılıç, A new adaptive support vector machine for diagnosis of diseases, Expert Syst. 31 (5) (2014) 389–397.

[9] M. Seera, C.P. Lim, A hybrid intelligent system for medical data classification, Expert Syst. Appl. 41 (5) (2014) 2239–2249.

[10] Y. Kazemi, S.A. Mirroshandel, A novel method for predicting kidney stone type using ensemble learning, Artif. Intell. Med. 84 (2018) 117–126.

[11] Shruti Ratnakar, K. Rajeswari, Rose Jacob, Prediction of heart disease using genetic algorithm for selection of optimal reduced set of attributes, Int. J. Adv. Comput. Eng. Netw. 1 (2) (2013) 51–55.

[12] Chul-Heui Lee, Seon-Hak Seo, Sang-Chul Choi, "Rule Discovery using hierarchical clasification structure with rough sets," IFSA World Congress and 20th NAFIPS International Conference, 2001.