ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 12| Issue: 3| March2026|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2026: 8.753 || ISI Value: 1.188

HEART DISEASE PREDICTION USING MACHINE LEARNING

Pon Mythili S[1], Mrs. P. Shanthi M.C.A., M.Phil., M.Sc., NET[2]
[1]III Department of Artificial Intelligence and Machine Learning, Dr.N. G.P. Arts and Science College
Coimbatore, Tamil Nadu, India
Guide, Assistant Professor

ABSTRACT

Early diagnosis of chronic medical conditions can lead to enhanced medical care and decreased mortality. Heart disease, diabetes, Parkinson's, kidney and liver disease impact millions of people around the globe. In the past, diagnosing decreasing deaths from chronic disease was performed through manual examinations of patient records and lab reports which could take a lot of time leading to patients sometimes being misdiagnosed or not receiving the right treatment for their illnesses in a timely manner. This paper presents a multi-disease prediction system developed through machine learning capable of predicting several diseases' chances at once by analyzing patient information such as age, sex/gender, Body Mass Index (BMI), blood pressure, cholesterol level, glucose level, physical habits (lifestyle factors) and genetic markers. The proposed model was developed using XGBoost and MultiOutputClassifier algorithms to improve the accuracy of disease prediction. Publicly available datasets were retrieved from Kaggle for training and assessing the performance of the model's performance measures (i.e., accuracy, precision, recall, F1 score) will be used thus validating the predictive nature of the system.

KEY WORDS: *Machine Learning, XGBoost, Multi-Disease Prediction System, Healthcare Analysis, Disease Risk Prediction System, Medical Data Analysis.*

1. INTRODUCTION

Chronic diseases are one of the leading causes of death worldwide and include significant illnesses such as heart disease, diabetes, Parkinson's disease, chronic kidney disease, and chronic liver disease. Chronic diseases greatly diminish the quality of life of those affected, while at the same time driving up the cost of healthcare. Early diagnosis of these diseases can help to prevent the progression of the disease, which decreases morbidity and mortality among patients. The standard approach to diagnosing chronic diseases is based on testing and manual assessment by healthcare providers. While these methods are often reliable forms of diagnosis, they can be time consuming and do not make the best use of the large quantities of data stored in electronic medical records.

As machine learning and data analytics have advanced, it is now possible to analyze large amounts of medical data in a more timely manner and to apply machine learning to assist providers in diagnosing chronic diseases based on analysis of the data. The overall objective of this project is to develop a system that uses machine learning algorithms to be able to identify patterns associated with multiple chronic diseases at once and to allow the user to predict the likelihood of developing one or more chronic diseases.

2. SYSTEM ANALYSIS

2.1 Current System

Current methods for predicting diseases in healthcare are typically based on standard methods of diagnosis or a very basic form of machine learning. Doctors usually review patient records for key factors (e.g., age, blood pressure, cholesterol level, ECG results) before making an assessment about whether or not the patient has a risk of developing a certain disease. In most cases, healthcare staff analyze this information manually leading to considerable amounts of time being spent performing these tasks.

Some methods of disease prediction which have been designed with the intention of utilizing machine learning algorithms exist; however, many are limited in their ability to predict only one particular disease and rely too heavily on standard classification type algorithms which may not be able to accurately represent complex data sets with regards to diagnosing medical conditions. Additionally, due to the amount of large amounts of healthcare data produced by hospitals, laboratories and wearable technologies; the current methodologies utilized to assess and process this data do not necessarily yield effective or efficient results.

The other major limitation with regard to current methodologies for predicting disease is that much of the work is done manually by healthcare staff; and therefore, there is little to no automation or integration with sophisticated analytical methodologies with which to process or assess data. Since all work is performed manually via healthcare staff; this results in an extended period of time before the diagnosis is completed which can lead to inconsistencies in decision making. Consequently, traditional methods for predicting disease are rarely capable of providing accurate or timely predictions which are necessary for the early identification of diseases and quality preventative healthcare.

Like other prediction systems, the one being developed will use patient health and lifestyle data to help identify who may be at risk for developing one or more chronic diseases. Unlike many prediction systems, this project's developed system will be able to predict multiple chronic diseases at the same time, using the same source of health and lifestyle data. The use of the XGBoost algorithm in this project is based on its historical

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 12| Issue: 3| March2026|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2026: 8.753 || ISI Value: 1.188

success as a classifier, in addition to its ability to efficiently handle large, complex data sets. The project will develop a system that will serve as a decision support system for healthcare providers in assisting with the early diagnosis of chronic disease(s), as well as supporting preventative health efforts.

2.2 Proposed System

The proposed system aims to develop a multi-disease predictive system, utilizing machine learning algorithms to predict multiple diseases by examining patient data. Patient's medical data will be utilized along with numerous health indicators to predict potential disease risks early on. It will predict heart disease, diabetes, Parkinson's, kidney, and liver diseases.

The system uses an XGBoost predictive model along with a MultiOutputClassifier. The XGBoost is an advanced machine learning algorithm that is well-known for high efficiency and accuracy when processing structured or tabular datasets. In addition, when the XGBoost is combined with a MultiOutputClassifier, one model can predict multiple disease types at once from one unified dataset.

This method increases the efficiency of disease prediction processes by automating the analysis of the data, therefore, reducing reliance on manual evaluation. Additionally, the proposed system can process large and complex patient datasets and find patterns in those datasets in order to generate accurate disease predictions. Therefore, the system will ultimately assist in providing early detection of potential risks to patients' health so that preventative measures can be taken by both patients and their caregivers prior to disease development.

3. SYSTEM REQUIREMENTS

3.1 Hardware Specifications

The system will require a reasonably spec'd computer in order to efficiently process and analyze medical datasets. A minimum recommendation would be to have an i3 or Ryzen 3 class processor, which should be suitable for data processing tasks. This means that an 8GB workforce will suffice to support machine learning functions and operate programs seamlessly. As a general rule of thumb, you would want at least 256GB of solid state (SSD) storage, which will allow for storage of datasets, computed models and project files. If you will be working with very large datasets or have computationally intensive tasks to perform, you could use a GPU to assist with computations.

3.2 Software Specifications

The requirements for building this development system are dependent on your development environment (DE); this could include Windows, Linux or MacOS. The programming language that has been chosen for working on this project is Python (3.8 or higher), as it is relatively easy to use, and contains a good degree of support for data science/machine learning development.

There will be a number of libraries to be utilized in order to execute the functions of the programme. The first library you will want to use when condensing data is pandas; the second will be to perform calculations using numpy. For performing machine learning functions, you are going to utilize scikit-learn; this library includes a host of resources for evaluating data that was used to create the model, preprocessing the data before feeding it into the model and splitting datasets.

4. METHODOLOGY OF THE SYSTEM

4.1 Dataset Collection

Our dataset for this study was obtained from publicly accessible datasets, such as Kaggle and other open medical repositories. The dataset contains health parameters for patients, including demographic characteristics (age, gender), body mass index (BMI), blood pressure, cholesterol level, smoking status, alcohol consumption, and physical activity level. Each instance of the dataset corresponds to a single patient and is classified as having one or more diseases (indicated by labels) and is sufficient to train the machine learning model.

4.2 Data Pre-Processing

Data pre-processing is completed before training the machine learning model to enhance the quality of data used for constructing the model. Pre-processing steps include the removal of missing data and duplicate records, which ensures accurate data usage. The categorical data (gender, smoking status, alcohol consumption) must also be converted into a numerical format. The final function applied to the data set is StandardScaler, which performs feature scaling, thereby allowing the data set to be distributed uniformly to enhance model performance.

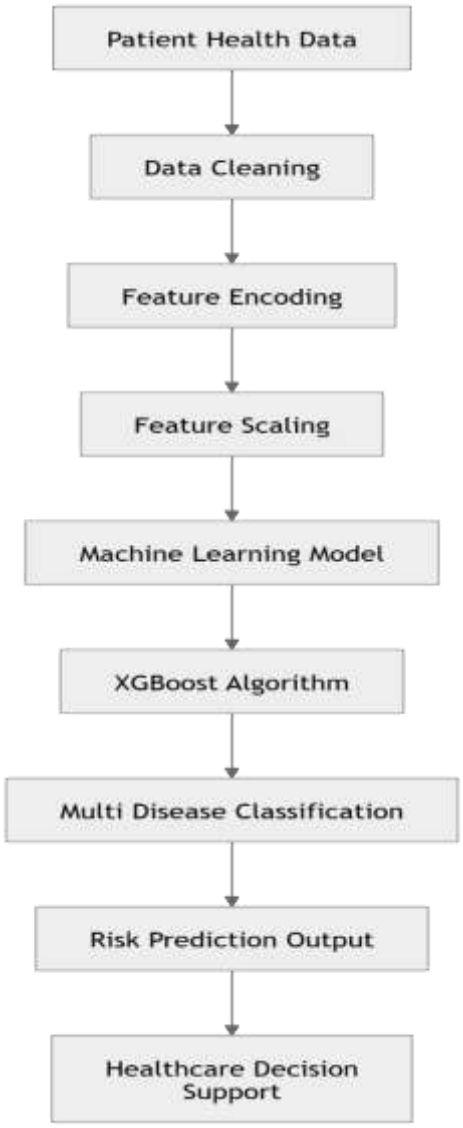
4.3 Model Training

The dataset is divided into two sets, 80% of the instances are used for training ((Train)) and 20% are used for testing ((Test)). The training is performed using the XGBoost algorithm to build multiple decision trees that are used to enhance the accuracy of the predictive model. To predict multiple diseases simultaneously using XGBoost, the MultiOutputClassifier is used in combination with the XGBoost algorithm.

4.4 Disease Prediction

Once the model has been trained, it uses the new patient information to generate predictions pertaining to disease risk. The new patient data will also be pre-processed prior to entering the trained model in order for the model to generate predictions. Each output prediction will indicate the predicted risk of development of various diseases based on patient input.

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 12| Issue: 3| March2026|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2026: 8.753 || ISI Value: 1.188



5. PERFORMANCE AND RESULTS EVALUATION ANALYSIS

There are multiple Metrics to evaluate performance for the Machine Learning Models, these include:

Accuracy- this metric is the overall number of correct predictions made by the model.

Precision- Is the number of prediction positive cases that were also actually positive cases.

Recall- Is the number of actual positive (truth example) cases that were accurately predicted as such by the model.

F1-Score- Is the metric for F1- Score which is a balance of both Precision and Recall.

The Experimental results produced by the model showed High accuracy and reliable prediction across multiple diseases.

6. CONCLUSION

This study showed that the multi-disease predictive system developed using machine learning and XGBoost could be used to analyze health, lifestyle, and genetics data of patients to predict their potential for developing a number of diseases. Results of this project demonstrated that using machine learning technologies can improve the timeliness and accuracy of identifying diseases at an earlier stage and assist healthcare practitioners with making more accurate diagnoses. The system created in this study is a method for improving disease prevention services that is both cost effective and scalable as a solution.

7. IMPROVEMENTS

Real-time health data from wearable devices (e.g., smart watches and fitness trackers) may be used in the future to enhance the system. Real-time patient health information will allow continuous patient health analysis and support more timely/detailed event forecasting. In addition, the system can be enhanced to contain prediction models for additional diseases, such as cancer and Alzheimer's disease. This will further enhance the tool's overall capabilities as a patient health monitor.

Additionally, advances in deep learning techniques potentially offer another source of improvement in the efficacy and efficiency of prediction estimates. In particular, the use of deep learning modelling affords a more powerful avenue for effectively analysing large and complex data sets, which can result in improved diagnosis support. Furthermore, the development of the system as a web or mobile application would facilitate easy access for users and healthcare providers from any location.

Lastly, if integrated with the electronic medical record systems used by hospitals, the system could retrieve and store patient medical records automatically. This could assist healthcare providers in analysing the patient's medical history, enabling them to provide prompt decisions regarding patient treatment, thus enhancing overall healthcare provision.

REFERENCE

1. *Detrano, R., et al., "The New Probability Algorithm for the Diagnosis of CAD: An International Application," Am. J. Cardiol. 1989.*
2. *Mitchell, T., Machine Learning, 1997.*
3. *Géron, A., Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow, 2019.*
4. *Chen, T., Guestrin, C., "XGBoost: A Scalable Tree Boosting System," 22nd ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, 2016.*
5. *World Health Organization (WHO), "CDC Statics for CVD's," WHO Official Website.*
6. *Kaggle, "Heart Disease Prediction Dataset," Kaggle Data Repository.*
7. *University of California; Irvine, UCI Machine Learning Repository, "Heart Disease Dataset."*
8. *Scikit-Learn Developers, "Machine Learning in Python," Official Doc.*
9. *XGboost Developers, "XGBoost Documentation," Official Guide.*
10. *Python Software Fundation, "Python Language Reference Manual."*