International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Predicting Diseases with Machine Learning

Bhoomika J Shetty[1], Ranjitha R[2]

[1,2]Lecturer, Department of computer science Sahyadri Science College, Shivamogga, Karnataka

Abstract:
Disease prediction is the estimation of the likelihood that an individual will develop a specific disease at some future time, based on a constellation of symptoms, laboratory tests, medical history, lifestyle factors and clinical markers. Early detection of disease may lead to optimal patient care, diagnosis, treatment, and ultimately prevention. The exponential growth of electronic health records, wearable healthcare devices, and large medical databases has made machine learning a promising means to support disease prediction and assist healthcare professionals in making correct clinical decisions.
Traditionally, a doctor diagnoses a disease by experience, carrying out clinical examinations and laboratory tests. This, while accepted and effective in many cases, can be a very time consuming, costly and highly skilled process. Therefore, the similarity in the symptoms of diseases can hamper patient health due to delayed clinical decision making to initiate the treatment, and also raise the cost of healthcare. As such, smart prediction systems for handling medical data are essential in providing reliable timely assistance through appropriate decisions and information to medical practitioners.
Machine learning refers to a data-driven approach that allows computers to discover hidden relationships and patterns in existing medical data, without being explicitly programmed for each individual condition. Machine learning algorithms can explore different healthcare information, such as patient demographics, symptoms, clinical measurements, diagnostic reports, and prior medical records to reveal complex trends that may be challenging to discern using traditional methods. The accessibility of large and various healthcare datasets has improved the capability of these models to generate accurate and reliable disease predictions.
Supervised machine learning algorithms have been widely used in disease prediction tasks because of its ability to classify and analyze medical data. Decision Tree algorithms are useful for clinical interpretation because they provide easily interpretable prediction rules based on patient characteristics. Random forest combine several decision tree is to improve prediction accuracy that are standard, which makes the prediction more accurate, stable and less prone to overfitting. The Support Vector Machine is useful in large scale medical data sets, because it can find the optimal boundary between various categories of diseases. Moreover, algorithms such as Logistic Regression, Naïve Bayes, K-Nearest Neighbors, Gradient Boosting, and Artificial Neural Networks have shown great potential in healthcare prediction applications. The efficiency of each algorithm is dependent on the dataset, complexity of the disease patterns and prediction goals.

Keywords: Artificial Intelligence, Machine Learning ,Naive Bayes, Support Vector Machine , Logistic Regression, Random Forest ,Gradient Boosting.

1. Introduction:

Prediction of diseases accurately and timely is crucial for good healthcare outcomes because it helps to

diagnose diseases early, formulate the right treatment strategies, and prevent any health complications. Early detection of diseases will help to reduce the cost of treatments, prolong patient life, and provide high-quality healthcare service. Due to fast developments in healthcare technology, large amounts of patient data have been collected from different sources such as demography, medical records, diagnostic data, and health monitoring devices. Analysis of these data can offer an opportunity to formulate intelligent systems that can assist healthcare providers to make decisions.

In Occurrence of a disease is not only affected by clinical factors, but also by external conditions, such as geographic location, environmental characteristics and individual lifestyle patterns. Climate conditions, environmental exposure, dietary habits and socioeconomic factors may be responsible for differences in disease prevalence from one region to another. Thus, by integrating contextual information with patient-specific medical data, the performance of disease prediction models can be improved. This comprehensive strategy including clinical and non-clinical factors can provide more reliable predictions and help to develop personalized healthcare solutions.

In normal health care delivery systems, people normally consult a general doctor once they start showing some symptoms. In such instances, the doctor performs initial checks and suggests that they consult a specialist depending on the condition they have. Despite being efficient, this system might need more time and energy since symptoms sometimes overlap many diseases. Disease prediction systems using machine learning techniques can assist in this process through examining the data of the patient to come up with probable health conditions according to their previously acquired knowledge from medical datasets.

With the rising popularity of health-related websites, people have started searching for health-related information and assessing themselves through such sites. But self-assessment using online resources may not yield accurate results due to the absence of any professional knowledge and the nature of symptom complexities of diseases. This is where disease prediction using machine learning comes into play, where there will be a preliminary prediction by such a system using the symptoms and health parameters provided by the user and at the same time motivating them to consult with healthcare experts for further confirmation and treatment.

The machine learning approach has proved highly beneficial in the application of medicine owing to its capability of finding patterns in big medical data sets. It has been found that classification algorithms can study different aspects of patients such as their age, gender, symptoms, vitals, past health record, and other health parameters that can be calculated with sensors, like blood pressure and heart rate, in order to find the probability of having some certain disease. Some commonly used classification algorithms include Logistic Regression, K-Nearest Neighbors, Naive Bayes, and Random Forest, Decision Tree.

From all of these methods, Random Forest stands out for its capability to deal with complicated data, minimize classification errors, and increase prediction accuracy due to ensemble learning. Just as that, other classifiers also play an important role in predicting diseases by finding correlations between input parameters and potential outcomes. In other words, with multiple patient parameters considered, the prediction may be improved.

The study involves creating an algorithm to predict the possible diseases based on the symptoms of the patients, demographics, physiological measures, and other information using machine learning techniques. For this purpose, different classifiers like Logistic Regression, Naive Bayes, K-Nearest Neighbor, Decision Tree, and Random Forest are used to compare the performance of the classifiers in disease detection. In addition, by predicting the right medical specialist for the disease identified, the patient may get proper treatment without delay.

2. Literature Survey

Malyadri, N., Raj, S., Choudhary, S., & Singhal, H. (2023) – Disease Prediction Using Machine Learning Malyadri, N., Raj, S., Choudhary, S., & Singhal, H. (2023) – Disease Prediction Based on Machine Learning Malyadri et al. (2023) suggested a methodology of using machine learning for disease prediction based on analysis of symptoms experienced by the patients. The study of Malyadri et al. (2023) examined various supervised machine learning algorithms like Decision Tree, Random Forest, and Support Vector Machine to identify the most appropriate one for disease prediction. According to the authors, the Random Forest model appeared to be the most accurate algorithm for predicting diseases due to its ability to process complicated data and avoid over fitting.

After analysing the importance of precise predictions of the models, Sabeena and Sruthika (2025) designed an intelligent model for predicting diseases, which aimed at enhancing the efficiency of the models through appropriate pre-processing of the data and feature selection. In the study, the importance of giving importance to the quality of input data was highlighted because the input data played an important role in the accuracy of machine learning models. Through proper use of features and classification, the efficiency of predictions could be increased.

Karachi, D., Penda Riya, A., Chandrakar, H., & Barapatre, O. P. (2024). Disease Prediction Using Machine Learning Kovachi et al. (2024) developed a machine learning model that would help in predicting different types of diseases based on the symptoms displayed by the patient. Different machine learning algorithms such as Naïve Bayes, Decision Tree and Random Forest were compared in order to find out which one is more efficient. As per experiments, it was found that results generated by using ensemble learning were more accurate as compared to classification algorithms. It was also stated by the authors that machine learning can help in improving healthcare services.

While a lot of studies have been carried out concerning the issue of predicting diseases based on their symptoms, Gaurav et al. (2023) extended the idea of disease prediction by introducing additional variables pertaining to the patients' age group, gender, Routine , and past medical history. Gaurav et al. argued that multiple dimensions of patient's data should be considered in order to predict diseases accurately rather than focusing on their symptoms alone. It was revealed that there were different machine learning techniques employed in this study and the best results were achieved by using ensemble learning technique.

Reddy, P. P., Babu, D. M., Kumar, H., & Sharma, S. (2021) – Machine Learning-Based Disease Prediction The authors Reddy et al. (2021) used machine learning algorithms to predict the disease of patients according to their symptoms. The processes of data pre-processing, feature extraction, and classification were performed to develop a predictive model for predicting diseases efficiently. They have revealed that machine learning methods can diagnose patients faster and more accurately. Such types of systems can act as a decision support system for healthcare professionals.

3. Proposed System

The same work seeks to build an intelligent and efficient disease prediction framework using machine learning technology that will help identify the possible diseases through analysis of the patient's health information. It follows a systematic process comprising requirement analysis, healthcare data gathering, pre-processing, exploratory analysis, modelling, performance evaluation, and optimization phases. All the phases play a major role in improving the accuracy and dependability of the resulting prediction model.

3.1 Requirement Analysis and Problem Statement

In the first phase of the proposed system, there is the need to analyse the healthcare problem and understand the objectives of the proposed prediction framework. Goal of the proposed system is to predict the potential diseases in a person by analysing the symptoms provided by the patients among other parameters related to their health status. At this stage, the required input variables, system requirements, and expected outcomes are analysed.

3.2 Data Acquisition and Pre-processing

Following the system requirements analysis, the next phase involves the acquisition of pertinent healthcare data needed in developing the machine learning model. Datasets that contain health-related information are acquired from trusted sources and may involve patients' symptoms, demographics, medical history, diagnoses, among other factors. Datasets could be available in different forms like database, spreadsheets, text files, or even health-care repositories. The availability of sufficient data is important since machine learning models require historic data in order to learn relationships and pattern from patients' symptoms to diseases.

However, healthcare data acquired in raw form is likely to have inconsistencies, incomplete data, redundant data, and even irrelevant data which may impact negatively on the accuracy of predictions made. Hence, there is need for data pre-processing in order to optimize the data acquired.

3.3 Data Pre-processing and Exploratory Analysis

When pre-processing phase of the process has been completed, exploratory data analysis is done to drive insight about the nature of the healthcare data that has been collected. The dataset is analyzed for identifying significant patterns, relationship among variables, and influencing factors for predicting diseases. Methods such as missing values treatment, outlier detection, removal of inconsistent data points, and feature selection are used for preparing quality data.

This allows for determining the significance of various patient features and ensures that only significant data is being fed into the machine learning algorithms. Data pre-processing becomes critical in improving the learning ability of the models.

3.4 Development and Training of Machine Learning Models

Here models is performed after the dataset preparation. In this step, dataset is split into test and train datasets. In the training dataset is utilized in teaching the machine learning model to recognize patterns and relations between patient’s features, symptoms, and possible diseases. Different classification models could be utilized, including Logistic Regression, Naïve Bayes, and Decision Tree, and Random Forest, K-Nearest Neighbors, models to determine the most appropriate model for disease identification.

In the training process, the algorithms discover hidden patterns from past medical data and generate predictive models that classify new cases of patients. Trained models are tested with unseen data to evaluate their predictive accuracy.

3.5 Evaluating the Model and Testing its Performance

Performance of the created machine learning models will be evaluated using the testing data set to ascertain how effective and reliable they are. There are various ways through which performance of a model could be analyzed such as accuracy, precision, recall among others. Through testing with new healthcare data, it would be possible to know if the system can generalize.

Results from the evaluation process are compared across the various algorithms to choose the one offering the best balance between accuracy and reliability.

3.6 System Enhancement and Model Optimization

The last step of the proposed system deals with optimization of the performance and efficiency of the disease prediction model. Various optimization techniques are used to make the model more efficient at performing predictions, which include choosing the best algorithm for the job, optimizing feature selection methods, adjusting the model parameters, and adding more healthcare data.

Consistent improvements in the quality of the data, the way features are represented, and the configuration of the model lead to the creation of a more accurate and efficient prediction system that will be able to deliver better disease identification performance and help both patients and healthcare professionals make well-informed healthcare decisions.

Overall, the suggested framework is expected to utilize data analysis and machine learning methods in developing an efficient disease prediction model. The framework seeks to make the disease detection process more effective by integrating healthcare data and classification techniques.

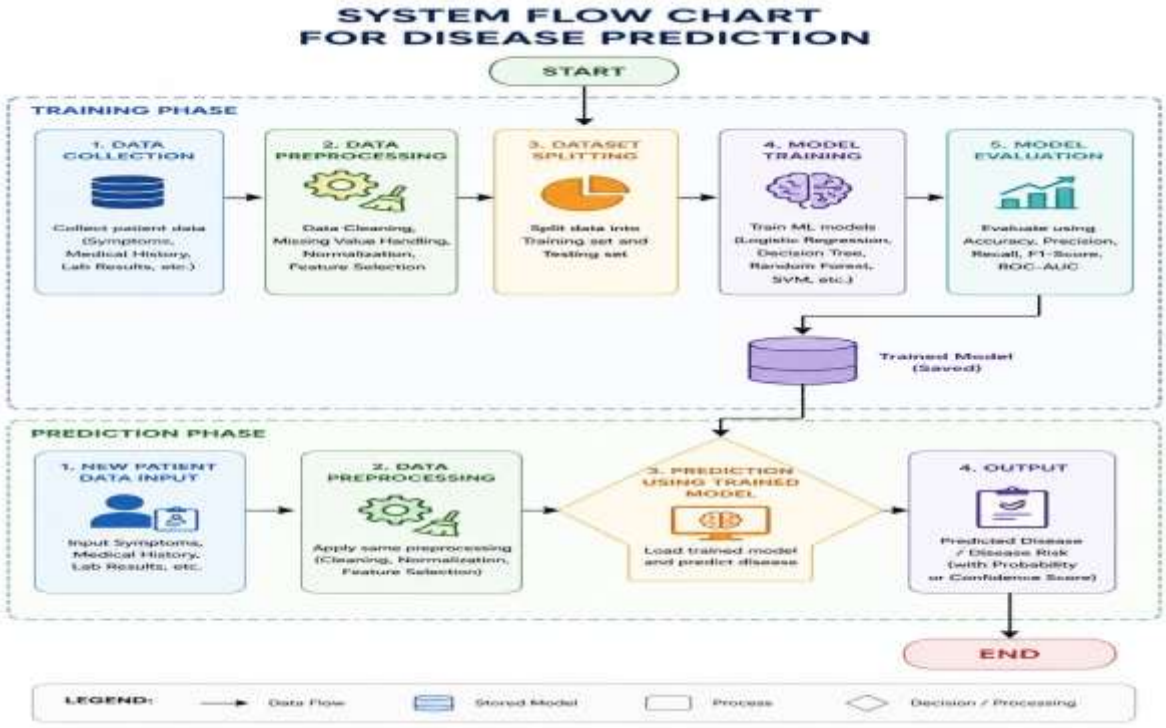

Figure 1: Flow chart

4. METHODOLOGY:

The suggested system for disease prediction is based on a systematic methodology using machine learning techniques consisting of various phases including acquiring patient data, pre-processing of data, analysing features, developing a model, evaluating performance, and predicting diseases. The most important goal of using the methodology is to examine various various component of patient data and design an effective model for predicting diseases. The overall methodology guarantees the conversion of raw health care data to useful knowledge through appropriate data processing and classification.

4.1 Patient Data Collection

Patient data forms the basic input in the proposed disease prediction system. The success of the prediction process depends mainly on the reliability of healthcare data. The proposed system takes into account different attributes of the patients, which include demographic data, health history, symptoms, lifestyle

and physiological attributes. Such attributes play an important role in understanding the link between patient attributes and diseases.

These data enable the machine learning algorithm to discover underlying patterns related to different diseases. For instance, data like age, gender, and symptoms can greatly help in predicting the probability of certain diseases.

Example:

Age = 45 years

Gender = Female

Symptoms = Fever and cough

Goal: To obtain necessary patient health data for disease prediction and analysis.

4.2 Data Collection for Healthcare

Data collection is an activity aimed at gathering all the required healthcare data that would be used to build and assess the predictive disease model. Sources of healthcare data can include electronic health records, clinical databases, laboratory test results, and generated data from patients.

The structured data consists of organized medical data such as patient demographics, laboratory test results, vital signs, and past medical records. The unstructured data may include symptoms stated by the patients and clinical findings documented by healthcare providers. Collection of diverse types of healthcare data enables a better understanding of the health status of patients and improves the efficiency of machine learning algorithms.

For privacy and security reasons, sensitive personal information such as patient names, identification number, and exact location is removed from the dataset. Healthcare data used for analysis include:

- Patient symptoms
- Medical history
- Laboratory test results
- Electronic Health Records (EHR)
- Clinical diagnosis reports

Goal: To collect meaningful healthcare data for development of the disease prediction model.

4.3 Data Pre-processing

Original healthcare databases may have incomplete, inconsistent or irrelevant data which may have an adverse effect on the performance of the machine learning algorithms. Thus, data pre-processing is performed to boost the consistency and quality of the data obtained from the medical sources before developing any prediction model.

In this step, several kinds of pre-processing activities are performed such as dealing with missing data, deleting duplicate data, resolving inconsistencies, cleaning data by eliminating unnecessary characters or space and formatting of data into required formats. The numeric data is normalized in order to ensure consistency among different measuring scales while categorical data is converted to numeric form.

The pre-processed database serves as the base of further analysis and machine learning model construction.

Goal: To convert original medical data into standardized and machine readable data.

4.4 Feature Extraction and Feature Selection

Here vital process that decide the success of a disease prediction system. Feature extraction is the process where raw data about patients is transformed into understandable features used in machine learning

models. Feature selection picks out the most relevant features having great influence on disease detection and removes the irrelevant or redundant ones.

Features in health care that should be considered for the purpose of prediction might include:

- Blood pressure
- Blood glucose
- Heart rate
- Cholesterol
- BMI
- Patient symptoms
- Information about age and gender

Feature selection minimizes the complexity of computations, increases the efficiency of machine learning and makes prediction more accurate because it allows the model to pay attention to the most influential features of disease occurrence.

Goal: To discover patient features that help in making disease predictions accurate.

4.5 Development of Machine Learning Model

Having completed data preparation and feature selection, the healthcare dataset prepared through the process is submitted to machine learning models for developing a predictive model. The primary focus at this stage would be on the identification of relationships between patient features and the disease condition based on historical data analysis.

The classification algorithms that are implemented for the purpose of comparison in order to select the best algorithm for predicting diseases are Logistic Regression, Random Forest, SVM Naïve Bayes, Decision Tree, and XGBoost.

4.5.1 Logistic Regression

This is an example of supervised classification model designed to predict the likelihood of developing certain diseases based on input patient features. This technique creates a relationship between independent variables (e.g., symptoms and medical parameters) and target classes (diseases). The simple structure and computational performance of Logistic Regression enable its widespread use in healthcare prediction algorithms.

4.5.2 Naïve Bayes

This is a probabilistic classification method that uses Bayes' theorem to calculate the likelihood of certain disease categories based on the analysis of patients' symptoms and health parameters. The technique is computationally fast and performs effectively even when working with extensive medical datasets with many attributes.

4.5.3 Decision Tree

This is a supervised classification model that separates input patient data using hierarchical tree structures. The model splits the dataset based on key features and creates decision rules that are easily understandable and interpretable. However, when used on complicated datasets, Decision Tree models may over fit.

4.5.4 Random Forest

This is an ensemble learning method designed to enhance prediction accuracy via combination of several Decision Trees. Each Decision Tree is trained on different data and feature subsets. The prediction outcome is generated by combining predictions obtained with several Decision Trees. As Random Forest

is a method suitable for complex healthcare data analysis, it is extensively applied in disease prediction tasks.

4.5.5 Support Vector Machine

This is a labelled based learning method that allows creating a decision boundary between different categories of diseases. This method is useful for high-dimensional data sets and enables producing classification results based on extracted meaningful features of healthcare information.

4.5.6 Extreme Gradient Boosting

This is an modern ensemble learning approach aimed at creation of more powerful classification model by combination of weak predictive models. This method involves the use of optimization and regularization techniques to increase accuracy and reduce overfitting.

The performance of all developed algorithms is evaluated to select the most reliable and accurate one for disease prediction.

4.6 Model Testing and Training

Once the algorithms are built, the dataset will be split for training and testing purposes. Training involves showing some patients' data to the algorithm in order for it to learn the relationship between the patient and disease features. Therefore, the algorithm will discover some relationships that can be useful in making predictions in the future.

New patients' data will be given to the model, and we will observe the accuracy of the algorithm in predicting the disease condition of the patients.

Goal: Train and test the classification model.

4.7 Performance Evaluation of Machine Learning Algorithm

Performance of the designed machine learning algorithm is measured through several performance metrics. The performance metrics assist in determining the effectiveness, efficiency, and classification capability of the particular machine learning algorithm.

The following are several performance metrics:

- **Accuracy:** Percentage of disease cases predicted correctly by the algorithm.
- **Precision:** Ratio of correct predictions of positive cases out of all predictions made by the algorithm.
- **Sensitivity (Recall):** Capability of the algorithm to correctly classify disease cases.
- **F1 Score:** Combination metric used to evaluate precision and recall.
- **ROC-AUC:** Classification of different disease classes by the algorithm.

Comparative analysis of the above mentioned performance metrics assists in determining the best machine learning algorithm for designing disease prediction system.4.8 Disease Prediction.

Once the machine learning algorithm has been trained and evaluated, it is used to predict the diseases from the new patient data inputs. The system will analyze the data about symptoms, demographic data, and medical parameters of the disease.

The prediction result could be:

- Disease present
- Disease absent

One among many other possible categories of disease when more than one disease is considered

The results will indicate the possibilities of diseases that the patient might be suffering from and help in making decisions for the clinicians.

Goal: Identify disease condition from the new patient health information.

4.9 Result Presentation

The last step of the methodology is the result presentation that should be easy for understanding and perception. The system shows the predicted disease as well as its confidence level or probability obtained from the machine learning algorithm.

Example:

Predicted Disease: Diabetes

Prediction Probability: 96%

Confidence Score: 0.96

With the help of confidence scores, it becomes easier for users and medical professionals to understand the reliability of prediction. Overall, the output will be beneficial in decision-making and will enhance usability of the disease prediction system.

In conclusion, the proposed methodology includes the combination of data processing in healthcare and machine learning approaches for creating an intelligent disease prediction system. Using patients' data and applying advanced classification methods, the system is able to detect diseases and make decisions in the field of healthcare.

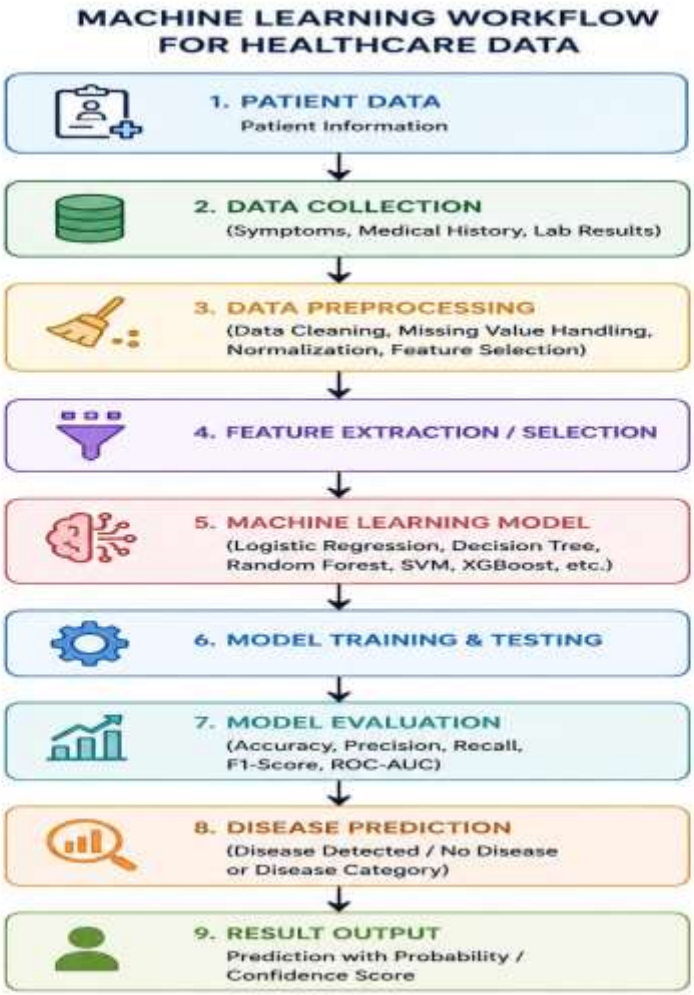

FIGURE: Work Flow Prediction

5. RESULT AND DISCUSSION:

The effectiveness of the proposed disease prediction system was investigated through the implementation and evaluation of several supervised machine learning techniques, such as logistic regression, naïve bayes, support vector machine, decision tree, random forest, and XGBoost. The health care database which was employed for constructing the models was partitioned into training and testing sets in order to investigate the learning capacity and prediction skills of each technique. In this regard, the training data set was utilized in finding the relations between patients features and diseases behaviour, while the testing data set was applied in evaluating the prediction capabilities of the models constructed using the training data.

Prior to implementing machine learning algorithms, the gathered healthcare data was pre-processed in order to make it more suitable for modelling. Such processes as data cleaning, which included such tasks as dealing with missing values, eliminating inconsistent records, filtering redundant data, and scaling numerical features, played an essential role in the process of enhancing models' performance. Furthermore, feature selection enabled the algorithm to focus only on the most important variables.

Performance Analysis of the Developed Models. The developed models were evaluated for their performances using different performance measures such as accuracy, precision, recall, and F1 Score. Performance analysis of these models gave us a better understanding of the classification capabilities of these models. Accuracy was used to evaluate the performance of these models in terms of correctness of predictions. Precision and Recall were used to test the efficiency of the models in predicting disease and non-disease cases.

The results show that the machine learning techniques can be applied successfully to study complicated healthcare data and detect patterns that correspond to particular diseases. The suggested system is quite unique compared to other traditional diagnostic approaches that require manual analysis, lab tests, and interpretation carried out by specialists. Even though the traditional approaches are still very important, machine learning prediction systems can help in analysing patients' data faster and give an indication about potential diseases.

In terms of machine learning algorithms used in the project, ensemble-based approaches like Random Forest and XGBoost proved to be efficient because of their high efficiency in predicting complex relationships in the medical datasets. In general, these algorithms consist of a number of learning models that enhance classification, prevent overfitting and increase reliability. For instance, Random Forest algorithm uses the power of several decision trees while XGBoost uses the optimized process of boosting. The rest of the algorithms (Logistic Regression, Naïve Bayes, Decision Tree, and SVM) also provided good classification results.

Generally, through the experimentation process, the study reveals that the disease prediction models using machine learning algorithms offer important help for early disease detection. Using symptoms of patients and medical data, the proposed model increases the effectiveness of health care analysis process and helps in making decisions about further diagnosis and treatment.

6. CONCLUSION:

Innovations brought about by using machine learning in medicine have led to the creation of intelligent systems that will be able to aid in predicting diseases and decision making in medicine. The disease prediction system suggested makes use of data related to patients such as symptoms, demographics, history, and clinical features of the patient in order to predict various diseases.

The proposed methodology consists of some key stages like gathering the information related to healthcare, its pre-processing, choosing the features, building the models, their training, testing, and performance assessment. Various classification methods like Logistic Regression, Naïve Bayes, Random Forest, Support Vector Machine, Decision Tree, and XGBoost have been considered and investigated in order to define their efficiency for prediction of diseases.

It is possible to state that the use of ensemble learning methods like Random Forest and XGBoost provides higher efficiency due to the presence of complicated patterns in healthcare datasets and the reduction of classification mistakes. Such algorithms are able to deal with large amounts of medical data successfully and increase the stability of prediction. However, other algorithms also have classification efficiency and may be effective depending on the dataset.

The suggested system will help healthcare professionals make earlier predictions about potential diseases, minimize the time spent on the initial analysis process, and help them make timely decisions regarding medical matters. Making early predictions about potential diseases can help to plan treatments better and cut down healthcare costs. The proposed system should be designed in such a way that it can act as a supplementary system to the clinical expertise of healthcare professionals.

Possible future developments for the proposed system can involve enhancing its capacity for predicting by using more extensive and diverse data related to the healthcare industry. The integration of the proposed system with the real-time EHR system, wearable health monitoring devices, and patient monitoring platforms will allow the system to offer more dynamic and personalized predictions. Moreover, employing advanced deep learning algorithms along with explainable artificial intelligence can increase the credibility of machine learning-based healthcare solutions.

Reference:

1. Malyadri, N., Raj, S., Choudhary, S., & Singhal, H. (2023). *Disease prediction using machine learning*. *International Journal of Creative Research Thoughts, 9*(11), 187–192.
2. Sabeena, S., & Sruthika, S. (2025). *Diseases prediction using machine learning*. *International Journal of Research Publication and Reviews, 6*(3), 9462–9476.
3. Kovachi, D., Pendariya, A., Chandrakar, H., & Barapatre, O. P. (2024). *Disease prediction using machine learning*. *International Journal for Research in Applied Science and Engineering Technology, 4*(12), 292–297. *(Please verify the page numbers, as the original "692–297" appears to be incorrect.)*
4. Gaurav, K., Kumar, A., Singh, P., Kumari, A., Kasar, M., & Suryawanshi, T. (2023). *Human disease prediction using machine learning techniques and real-life parameters*. *Journal of Survey in Fisheries Sciences, 36*(6), 1092–1098.
5. Reddy, P. P., Babu, D. M., Kumar, H., & Sharma, S. (2021). *Disease prediction using machine learning*. *International Journal of Engineering Research & Technology, 9*(6), 205–208.