ISSN: 2350-0328
International Journal of AdvancedResearch in Science, Engineering and Technology
Vol. 12, Issue 5, May 2025

Predicting Diseases with Machine Learning Algorithms

Sumit Vishwanath Wakode, Prof. Minakshi Ramteke

P.G. Student, Department of Computer Science, VMIT, Nagpur, India
Assistant Professor, Department of Computer Science, VMIT, Nagpur, India

ABSTRACT: It is necessary to analyze any health-related issue promptly and accurately in order to prevent and treat illnesses Using the traditional method may not provide an accurate diagnosis of a serious illness. A medical diagnosis system that predicts any disease using machine learning (ML) algorithms can provide a better diagnosis than the conventional method. A variety of machine learning algorithm were employed in the development of the disease prediction system. In the processed dataset, over 230 diseases were included. The diagnosis system is used to determine the disease that a person may have based on their age, gender, and symptoms. The results of the weighted KNN algorithm were the best of all the algorithms. A prediction accuracy of 93.5 percent was achieved by the weighted KNN algorithm. As a doctor for early disease detection, our diagnosis model can guarantee prompt treatment and lifesaving.

I. INTRODUCTION

Among the most crucial facets of human existence and the economy are healthcare and medicine. Significant changes have occurred between the world we live in now and the one that existed just a few weeks ago. Everything has changed and become graphic. In this scenario, where everything has gone virtual, the medical staff is working tirelessly to save certified doctors who choose to practice virtually via phone and video consultations rather than in person are referred to because they can perform tasks more rapidly and precisely when human error is eliminated. Without human error, a disease predictor-also referred to as virtual doctor-can forecast any patient's condition. A disease predictor can also be helpful in situations like-Covid-19 and Ebola because it can identify a person's illness without coming into contact with them. Although virtual doctor models exist, they fall short of the required accuracy standards due to their failure to consider all relevant factors. Developing a rang of models was the primary goal in order to identify which one generates the most accurate predictions. Even though the scale and complexity of ML projects vary, they all have the same general structure.an assortment of rule-based technologies.

A. DESCRIPTION OF THE PROBLEM.

The goal of disease prediction using machine learning algorithms was to minimize healthcare expenses as much as possible. A number of features in the health prediction system are still undeveloped. Therefore, even though we live in a technologically advanced world, it is useless if we cannot use it effectively and appropriately. Research is being done on health prediction systems to address this. Numerous applications make use of various technologies. This project demonstrates how to combine the two technologies to produce an effective outcome.

B. PROPOSITOS.

1. The Proposed Topic Scope:

- Incomplete quality of medical data lowers the analysis's accuracy.
- Furthermore, distinct regional traits are displayed by various regions.
- But those previous studies primarily focused on structured data.
- Semi-structured and unstructured data cannot be handled properly.
- In the proposed system, both structured and unstructured data will be taken into account.

2. The Proposed Topic Scope:

- Our aim is to develop a tool that will help consumers and professional identify and select diseases.

ISSN: 2350-0328
International Journal of AdvancedResearch in Science, Engineering and Technology
Vol. 12, Issue 5, May 2025

- Early disease detection will be very helpful to healthcare organizations like hospitals. We can add more illnesses to the current system in the future. To lower the death rate, we can attempt to increase prediction accuracy.

II. SYSTEM REQUIREMENTS

A. HARDWARE REQUIREMENTS

- •Processor: Core i3/i5/i7
- • RAM: 2-4GB
- • HDD: 500 GB

B. SOFTWARE REQUIREMENTS

- • Platform: Windows Xp/7/8/10/11
- • Coding Language: Python

III. SYSTEM ANALYSIS

Our system predicts a patient's disease based on their symptoms using algorithms and other tools. These symptoms are then contrasted with the system's previously accessible dataset. The machine learning system for such as multiple disease prediction that has been suggested is this one. The patient's condition will be compared with those datasets to determine the exact percu for later use, and then the user chooses the features by selecting or entering the various symptoms. Classifying those data sets is then done using machine learning algorithms such as logistic regression. The recommendation model receives the data after which it presents the system's risk analysis and provides a probability estimation of the system that illustrates the different probabilities, including the system's behaviour after a predictions are made. Additionally based on the patient's symptoms and final results, it suggests things for them to use and things they shouldn't use from the datasets that are provided and the final results. It makes predictions about probable diseases using data sets such as COVID-19, chronic kidney disease, and heart disease, and heart disease. To our knowledge, no existing research focuses on medical big data analytics.

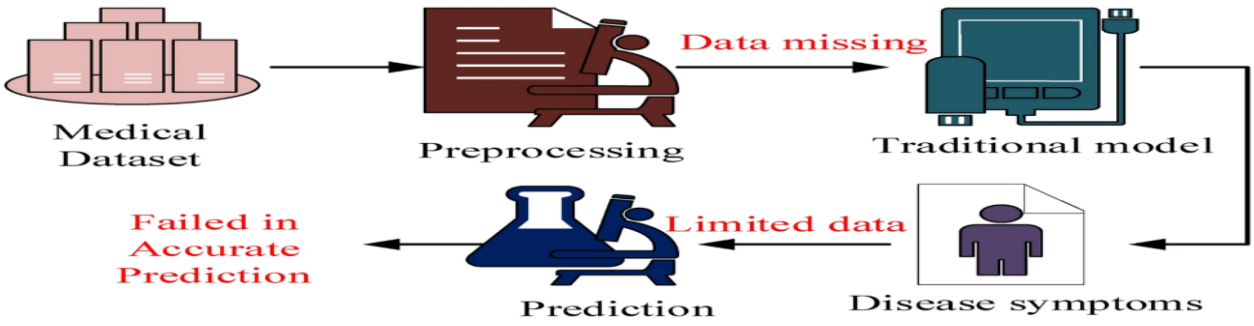

Fig. 1 Proposed system for disease prediction

Fig.1 A system for predicting diseases is suggested. The doctor may not always be available when needed. Nonetheless, given the current circumstances, this prediction system may be applied whenever necessary. By entering the person's age, gender, and symptoms, the ML model can be further processed. The machine learning model uses the current input to train and test the algorithm that generates the predicted disease after the initial data processing.

ISSN: 2350-0328
International Journal of AdvancedResearch in Science, Engineering and Technology
Vol. 12, Issue 5, May 2025

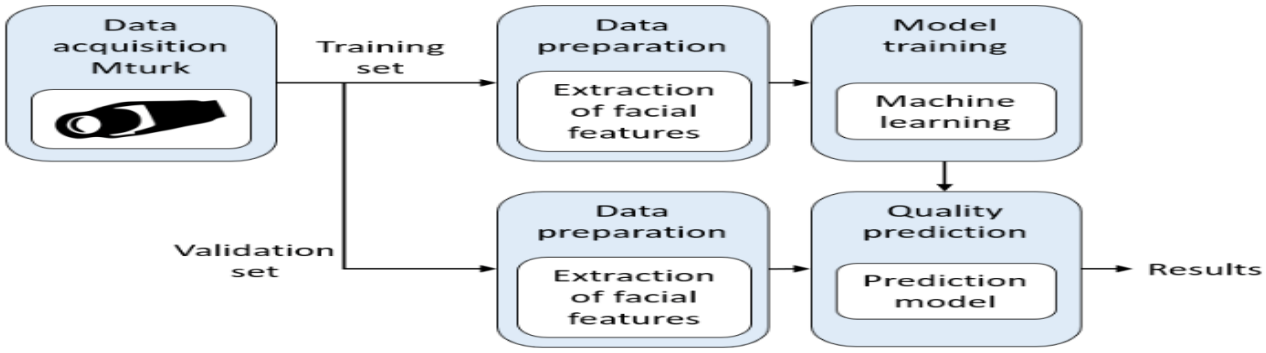

Fig. 2 Proposed system flow diagram during training

Fig 2 During training, the system flow diagrams were suggested. The pre-processed dataset, which was then fed into different machine learning algorithms to predict the disease, consisted of an individual's age, gender, and symptoms. Several machine learning models were used, included RUS Boosted trees, Fine, Medium, and Coarse Decision tress. Weighted KNM, Subspace KNN, Gaussian Naïve Bayes, Kernel Naïve Bayes, and Fine, Medium, and Coarse KNN. The processing model receives inputs such as age, gender, and the symptoms of the disease.

IV. WORKING OF THE ML MODEL

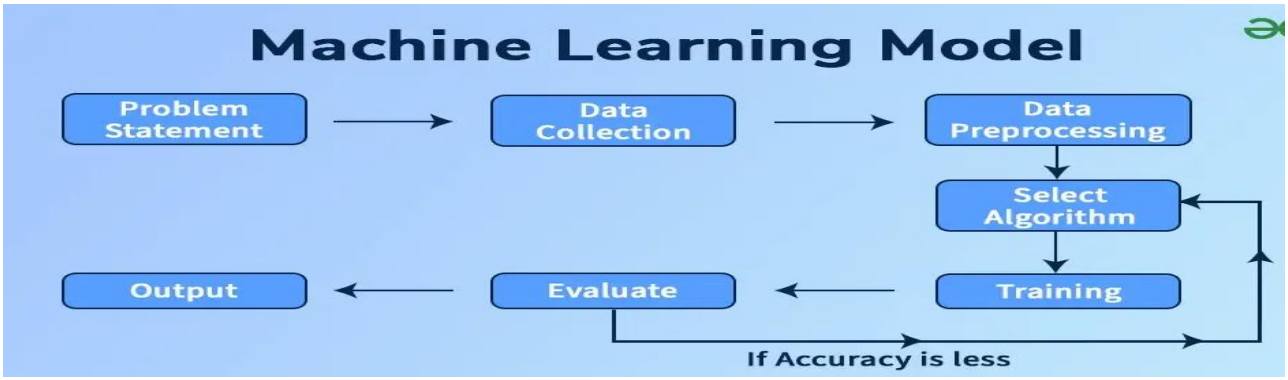

Fig. 3 Functioning of the ML models

Fig. 3. How the ML models function, Age, Gender, and symptoms were used as input factors to divide the datasets into inputs and outputs, which stood in for the diseases. At random, the available data was split into train and test sets. Following that, these sets were encoded and put through more training with different algorithms. After that, the algorithms asses the training set and predict the values, assessing the precision of different machine learning algorithms. The disease was the result of decoding the predicted values.

ISSN: 2350-0328
International Journal of AdvancedResearch in Science, Engineering and Technology
Vol. 12, Issue 5, May 2025

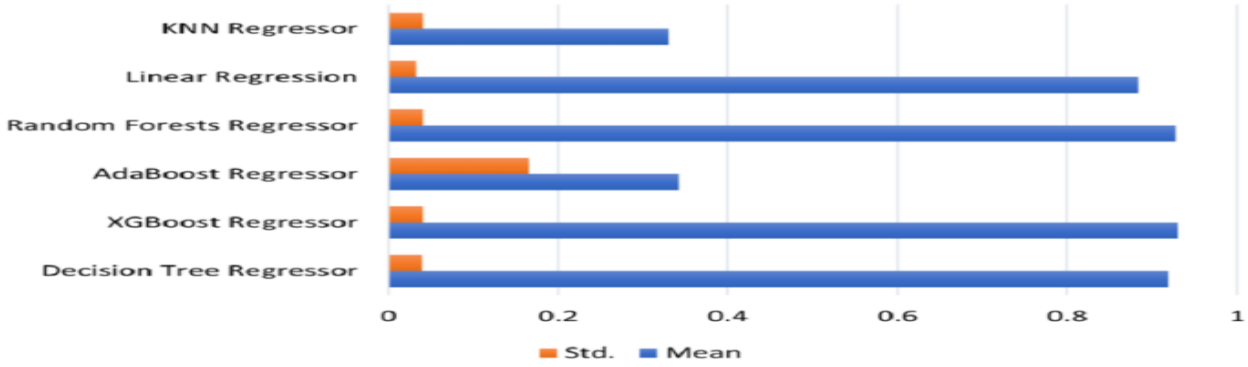

Fig .4 Accuracy values of different ML models

KNN stands for K-nearest neighbors:
The KNN algorithm is a type of supervised machine learning algorithm. It only calculated the distance between each new Data point and each other training data point. The distance could be Euclidean or Manhattan. Next, it selects the K data Points-where K can be any integer- that are closest to it. After that, the data point is assigned to the class to which the majority of K data points belong.

Weighted KNN:
Within the category of supervised machine learning algorithms is the KNN algorithm. It only computed the separation between every training data point and every new data point. Manhattan or Euclidean distances are both possible. It then chooses the K nearest data points, where K can be any integer. Following that, the data point is allocated to the class that the most K data points fall into.

Fine, Medium, and Coarse KNN:
Assigning integer value to K for Fine, Medium, and Coarse KNN is necessary in order to calculate the distance. Therefore, we set K to a low value in our fine KNN model, meaning that it uses about one neighbour to make predictions. In a similar vein, the medium KNN model uses about 10 neighbours while the coarse KNN use 100. The wide variation in accuracy percentages was also caused by the distinct neighbours of each model. When it came to accuracy, our fine KNN model beat the order two models, but the coarse KNN yielded a low prediction value for the data (regardless of the hypothesis).

Naive Bayes
The Bayes probability theorem serves as the foundation for this machine learning algorithm for classification problems. Using high dimensional training data sets for text classification is the main application for this. To determine the posterior probability, we applied the Bayes theorem, which is defined as follows:

$$\mathbf{P(h|d)} = \frac{\boldsymbol{P}(\boldsymbol{d}|\boldsymbol{h})\cdot\boldsymbol{P}(\boldsymbol{h})}{\boldsymbol{P}(\boldsymbol{d})}$$

is the probability of hypothesis h given the data ddot The probability of data d given that hypothesis h was true is denoted by P(d|h). The likelihood that hypotheses h is correct (regardless of the data) is denoted by P(h). The prior probability of h is what this is considered. P(d) is the data's probability, independent of the hypothesis.

Gaussian Naive Bayes:
The Gaussian version of the Native Bayes algorithm follows the same methodology. It is necessary to have a dataset with all continuous features for Gaussian Naïve Bayes and a categorical dataset for Naïve Bayes. This model had to be applied since our dataset's continuous features included symptoms, age, and gender. The accuracy rate for this model was not very high.

Kernel Naive Bayes
Our dataset included age and other numerical attributes, so we used Kernel Naïve Bayes to predict the medications. Similar steps are taken by this algorithm and the Native Bayes algorithm. The primary benefit of this algorithm is the

ISSN: 2350-0328
International Journal of AdvancedResearch in Science, Engineering and Technology
Vol. 12, Issue 5, May 2025

nonparametric is provides. If there is no prior information regarding whether the dataset being uses is parametric, this model may yield more accurate results. The Gaussian Naïve Bayes model and this model yielded results that were almost identical.

Decision trees
Within the family of supervised learning algorithms, the decision trees algorithm is 4.7. Its applications include classification and regression. For prediction, the tree diagram approach is applied at the top of the decision tree. It splits once in the dominant input feature after having a root node. Once all of the inputs have been added, these processes continue until the final node has weights; the input is then categorized based on these weights. In a coarse trees, there can be up to four splits per node. In a medium tree, on the other hand, node can have up to 20 splits. It is possible for a node in a fine tree to splits up to 100.

SubSpace KNN
Similar to bagging, the SubSpace KNN method substitutes features and randomly samples them. As a result, individual students pay less attention to features that appear to be highly descriptive or predictive within the training set but less predictive for points outside of it. Random subspaces are therefore a preferred choice for issues where the number of features is noticeably larger than the number of training points.

RUSBoost algorithm
To get accurate results, the RUSBoost algorithm needed our dataP to be properly trained. Using the RUSBoost algorithm, the trained data set derived from the skewed data set performs better. An algorithm called RUSBoost blends boosting and data sampling. This algorithm is among the most effective approaches.

V. ADVANTAGES AND DISADVANTAGES

A. ADVANTAGES

Proactive Disease Detection: Machine learning, which unveils patterns and deviations in large datasets, enables the early identification of diseases before their symptoms become apparent.

Accuracy: ML algorithms can analyze complex data more accurately than humans, leading to more reliable predictions.

Personalized medicine: By utilizing the unique health data of each patient, machine learning can develop treatment regimens that are more effective.

Earlh Intervention: Medical personnel can act quickly and potentially save lives by employing predictive models.

Cost-Efficiency: Early detection and prevention can reduce healthcare costs by preventing expensive treatments associated with advanced disease stages.

B. DISADVANTAGES

Prediction accuracy: Prediction accuracy is significantly impacted by the quality of the input data, which may contain biases or errors.

Privacy Concerns: Collecting and sharing personal health data for machine learning may raise privacy concerns if done incorrectly.

Overfitting: Machine learning model's propensity to overdraft to training data impairs their capacity to generalize to new contexts.

ISSN: 2350-0328
International Journal of AdvancedResearch in Science, Engineering and Technology
Vol. 12, Issue 5, May 2025

Resource-intensive: The deployment and maintenance of machine learning systems in the healthcare sector require infrastructure and trained personnel.

Ethical Concerns: Choices based on machine learning predictions could lead to ethical dilemmas, such as biased outcomes or discrimination against specific groups.

VI. CONCLUSION

The manuscript described how to predict a patient's disease based on their age, gender, and symptoms. The Weighted KNN model predicts diseases with the highest accuracy of 93% when using the previously mentioned factors. Results from almost all ML models were accurate. Some models had a low accuracy rate and were unable to predict the disease because they depended on the parameters. Once the disease was predicted, it would be easy to manage the medication resources required for treatment. This model would help lower the costs related to treating the illness while also enhancing the healing process. Machine learning has the potential to revolutionize disease prediction and improve patient outcomes. Examining the most popular algorithms and themes in diseases prediction research can greatly improve public health and healthcare. Let's work together to advance this exciting field and make people's lives better.

REFERENCES

[1]. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3661426
[2]. https://ijcrt.org/papers/ijcrt2105229.pdf
[3]. https://www.researchgate.net/publication/357449131_the_prediction_of_disease_using_machine_learning
[4]. https://nevonprojects.com/multiple-disease-prediction-system-using-machine-learning/
[5]. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3661426#:~:text=We%20have%20designed%20a%20disease,individual%20might%20be%20suffering%20from.