International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 05 | May - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

HEALTH PREDICTION USING MACHINE LEARNING

SAURABH MISHRA, SANSKAR KUMAR, MR. SHRAVAN KUMAR YADAV

B. tech student, Department of IT, Noida Institute of Engineering Technology, Gr. Noida
B. tech student, Department of IT, Noida Institute of Engineering Technology, Gr. Noida
Assistant professor, Department of IT, Noida Institute of Engineering Technology, Gr. Noida

---

Abstract - Machine learning techniques have transformed healthcare by enabling precise and timely disease prediction. The capacity to forecast multiple diseases simultaneously can greatly enhance early diagnosis and treatment, leading to improved patient outcomes and lower healthcare expenses. This research paper delves into the use of machine learning algorithms for predicting various diseases, highlighting their advantages, challenges, and prospects. It provides a comprehensive overview of different machine learning models and the data sources frequently employed in disease prediction. Furthermore, it emphasises the importance of feature selection, model evaluation, and the integration of diverse data types to improve prediction accuracy. The findings underscore the significant potential of machine learning in predicting multiple diseases and its impact on public health. Specifically, the study demonstrates the application of a machine learning model to determine if an individual is affected by certain diseases. This model is trained using sample data to enhance its predictive capabilities.

***Key Words*:** Disease Prediction, Disease data, Machine Learning.

INTRODUCTION

Machine Learning is a field that leverages historical data for predictive purposes. It involves developing computer systems that learn from data and experiences. Machine learning algorithms operate in two main stages: training and testing. Over the past few decades, machine learning technology has faced challenges in predicting diseases based on patient symptoms and medical histories. However, this technology has the potential to address healthcare issues effectively. By applying comprehensive machine learning concepts, we can monitor patient health more efficiently.

Machine learning models enable the rapid cleaning and processing of data, facilitating quicker results. This system supports doctors in making informed decisions about patient diagnoses and treatments, thereby enhancing healthcare services. Introducing machine learning into the medical field, particularly healthcare, is a prime example of its application. To improve accuracy with large datasets, existing work focuses on unstructured or textual data. For disease prediction, current efforts involve linear models, SK learn, and Decision Tree algorithms.

Machine learning is a field that utilizes historical data for making predictions. Using this approach will enable the creation of an intuitive user interface with the ability to predict multiple diseases with high accuracy using a single application. Accurate disease prediction using machine learning models can enable early interventions, personalised treatment plans, and targeted disease management strategies. This technology can assist healthcare providers in making well-informed decisions, improving patient care, and optimising resource allocation within healthcare systems. Additionally, it shows promise for population-level disease surveillance, helping public health authorities to detect outbreaks and implement preventive measures swiftly. This research contributes to the expanding body of literature on machine learning-based disease prediction, specifically highlighting the use of Decision Tree for predicting multiple diseases.

Technologies: Machine Learning using Python, Html Bootstrap flask for Front-End Development

Platforms: Visual Studio code, Google Collab for machine learning

Libraries: NumPy, Pandas, Scikit-learn

1. LITERATURE SURVEY (RELATED WORK)

Machine learning and artificial intelligence have become crucial components in various industries, including healthcare. Predictive models based on machine learning algorithms can accurately and quickly detect diseases, enabling doctors to provide better treatment and care for patients. Your project to detect multiple diseases, such as heart disease, liver disease, and diabetes, using machine learning algorithms is an excellent initiative. Algorithms like Random Forest and K-Nearest Neighbours (KNN) can help achieve high accuracy and enhance the predictive model's effectiveness. However, it's important to acknowledge that machine learning models are not flawless and have limitations. Therefore, validating the model's accuracy with real-world data and having a medical expert review the results is essential to ensure patient safety. Overall, incorporating machine learning and artificial intelligence in the medical field holds significant potential for advancing healthcare.

This system could enhance the efficiency and accuracy of disease prediction, aiding doctors in delivering better patient care. By employing machine learning algorithms and TensorFlow, you can train models to analyse multiple diseases simultaneously.

The Flask API can create a web service to receive user inputs, such as disease parameters and names, and then invoke the corresponding model to predict the disease status. The integration of machine learning and the Flask API in this

system offers several benefits, including faster and more accurate disease prediction, early warnings of potential health risks, and improved patient outcomes. Furthermore, this system can be expanded to include additional diseases in the future, increasing its utility and effectiveness. Overall, this approach has the potential to revolutionize disease diagnosis and treatment, saving countless lives through early detection and timely intervention.

The use of computer-based technology in healthcare has led to the accumulation of vast amounts of electronic data, complicating the analysis of symptoms and early disease detection for medical personnel. However, supervised machine learning algorithms have shown great promise in outperforming traditional diagnostic methods and assisting medical professionals in early identification of high-risk conditions. This literature review examined trends in using supervised machine learning models for disease identification, focusing on algorithms such as Naive Bayes, Decision Trees, and K-Nearest Neighbours. The findings indicate that support vector machines excel at detecting Parkinson's disease and kidney disorders, while logistic regression is used for heart disease prediction. High-accuracy predictions for breast diseases and common diseases were achieved using Random Forest and convolutional neural networks, respectively.

The integration of machine learning (ML) and artificial intelligence (AI) into healthcare has dramatically improved disease prediction, enabling more precise and timely diagnoses. Numerous studies have examined the efficacy of various ML algorithms in predicting diseases, showcasing their substantial potential to enhance patient care and outcomes.

In a notable study, K.M. Al-Aidaroos, A.A. Bakar, and Z. Othman compared the Naive Bayes (NB) algorithm with five other classifiers: Logistic Regression (LR), K-Star (K*), Decision Tree (DT), Neural Network (NN), and zero. Utilising 15 real-world medical datasets from the UCI machine learning repository, their research found that NB outperformed the other algorithms in 8 out of 15 datasets, underscoring its superior predictive accuracy in numerous medical applications (Asuncion and Newman, 2007).

Another significant study by Darcy A. Davis and colleagues developed the CARE system to predict potential disease risks. CARE uses a patient's medical history and ICD-9-CM codes, employing collective filtering and clustering techniques to predict disease risks based on the medical histories of similar patients. The researchers further enhanced this system with ICARE, an iterative version incorporating ensemble methods for improved performance. These systems can predict a wide range of medical conditions in a single run, providing early warnings for numerous diseases, which is crucial for preventive healthcare and early intervention.

Jyoti Soni, Ujma Ansari, Dipesh Sharma, and Sunita Soni reviewed existing data mining techniques for heart disease prediction. Their comparative analysis revealed that Decision Trees often provided the best performance, although Bayesian classification occasionally achieved similar accuracy. Other methods like KNN, Neural Networks, and Clustering-based classification generally showed lower performance, highlighting the variability in effectiveness across different algorithms.

Shadab Adam Pattekari and Asma Parveen focused on heart disease prediction using the Decision Tree algorithm. They found that patient-provided data, when compared against a prequalified set of values, could effectively predict heart disease. Similarly, M.A. Nishara Banu and B. Gomathy explored heart-related problems using various data mining techniques, such as association rule mining, grouping, and clustering. Their study emphasised the utility of decision trees in illustrating potential outcomes based on factors like age, sex, smoking habits, and other health indicators. They used K-means clustering to analyse dataset patterns, enhancing the accuracy of predictions by grouping data into clusters.

Collectively, these studies highlight the diverse applications and effectiveness of machine learning algorithms in disease prediction. They demonstrate the potential of these technologies to transform healthcare by providing early and accurate disease diagnoses, enabling timely and effective interventions. However, they also emphasise the importance of selecting the appropriate algorithm based on specific medical datasets and conditions to achieve optimal results. The ongoing advancements in machine learning and AI promise further improvements in predictive accuracy and clinical utility, potentially saving countless lives through early detection and treatment.

2. PROBLEM STATEMENT

Research indicates that most machine learning models developed for healthcare analysis focus on a single disease. For instance, one model may analyse liver issues, another may focus on cancer, and a third on lung problems. Consequently, individuals need to consult multiple online resources to obtain accurate predictions for various illnesses.

There is no well-defined process for predicting multiple diseases through a single analysis. Additionally, poor accuracy in some models can have serious implications for patient health. Companies seeking to evaluate their patients' medical records face increased time and financial costs due to the need to implement various models. Furthermore, some existing systems consider too few parameters, potentially leading to inaccurate results.

3. EXISTING SYSTEM

Despite advancements in computing, doctors still heavily rely on technology for various tasks such as surgical procedures and x-ray imaging. However, this technology has not fully met the medical field's demands. Doctors' knowledge and experience remain crucial due to a range of influencing factors, including medical records, weather conditions, the atmosphere, blood pressure, and other variables. While numerous variables need to be considered to understand the entire process, no model has successfully analysed all of them. To address this issue, medical decision support systems should be utilised. These systems can assist doctors in making accurate decisions.

We propose applying machine learning to manage comprehensive hospital data. Machine learning technology enables the creation of models that quickly analyse data and deliver faster results. By leveraging machine learning, doctors can make more informed decisions regarding patient diagnoses and treatment options, thereby improving patient healthcare services. The application of machine learning in healthcare exemplifies its significant potential in enhancing medical outcomes.

4. PROPOSED SYSTEM

We propose a system featuring a simple, cost-effective, and elegant user interface that is also highly time efficient. This system aims to bridge the gap between doctors and patients, facilitating both parties in achieving their healthcare goals. The system predicts diseases based on user-reported symptoms. Users will input five symptoms, which will be analysed using algorithms such as Decision Tree, Random Forest, Naïve Bayes, and KNN to ensure accurate predictions.

Our approach involves integrating and analysing diverse datasets from a broad population, enhancing the system's accuracy and reliability. This, in turn, helps build patient trust in the system's predictions. Additionally, the system will maintain a database to store user input and the corresponding predicted diseases, which can be referenced for future treatments. This feature will significantly contribute to more efficient health management and improved user satisfaction.

By merging and exploring extensive datasets, our system continuously improves in accuracy, offering reliable predictions and fostering patient confidence. This comprehensive database ensures that health management is simplified, and patient satisfaction is elevated.

5. METHODOLOGY

Our project focuses on predicting multiple diseases based on symptoms entered by patientsDefining the problem statement is the first step. Next, we prepare the dataset for analysis. We then visualise the data using scatter plots, distribution graphs, and other tools to identify anomalies and missing values, ensuring the dataset is ready for prediction.

The core feature of our project is the application of machine learning algorithms such as Decision Tree, Random Forest, Naïve Bayes, and KNN to achieve accurate disease predictions. This early prediction capability aims to significantly enhance patient care.

We have implemented this model using Python to execute our machine learning algorithms. Additionally, we have developed an elegant graphical user interface (GUI) to facilitate user interaction with the system.

To evaluate performance, we calculate the following metrics: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). These metrics allow us to derive four key measurements: recall, precision, and accuracy. True positives represent the number of correctly predicted positive results, true negatives represent the number of correctly predicted negative results, false positives represent the number of incorrectly predicted positive results, and false negatives represent the number of incorrectly predicted negative results.

The flow chart of the methodology is given below:

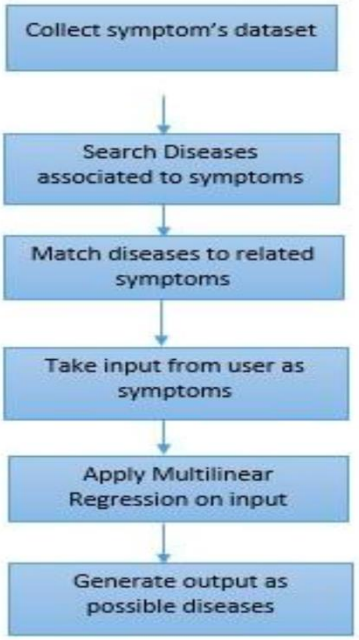


Fig 1: Flow chart of proposed method

Accuracy-:

$$\frac{TruePositive+TrueNegative}{TruePositive+TrueNegative+FalsePositive+FalseNegative}$$

$$\text{Precision} = \frac{TruePositive}{TruePositive+FalsePositive}$$

$$\text{Recall} = \frac{TruePositive}{TruePositive+FalseNegative}$$

$$\text{F1-Measure} = \frac{2\times precision\times recall}{precision+recall}$$

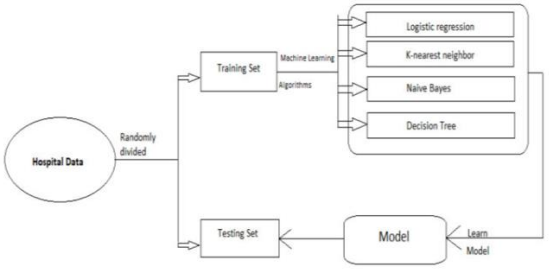


Fig -1: System Architecture

6. ALGORITHM TECHNIQUES

6.1 Decision Tree

The type of decision tree utilised in this research is the gain ratio decision tree. This approach is based on the concept of entropy (information gain), which aims to select the splitting attribute that minimises entropy, thereby maximising the information gain. Information gain represents the difference between the original information content and the amount of information needed after the split. The attributes are ranked according to their information gains, with the highest-ranked features being chosen as potential attributes for the classifier.

To determine the optimal splitting attribute for the decision tree, one must compute the information gain for each attribute and then select the attribute that provides the highest information gain. The formula used to calculate the information gain for each attribute is as follows:

$$E = \sum_{i=1}^{K} P_i \ log_2 \ \ P_i$$

Where *k* represents the number of classes of the target attributes and *Pi* is the probability of class *i,* calculated by dividing the number of occurrences of class *i* by the total number of instances. To mitigate the bias introduced using information gain, a variant called gain ratio was developed by Australian academic Ross Quinlan. Information gain tends to favour attributes with many values. The gain ratio adjusts the information gain for each attribute to account for the attribute value's breadth and uniformity.
Decision Trees are a supervised learning method used for both regression and classification. To estimate the value of the target variable, they acquire basic decision rules derived from the data attributes. Various decision tree algorithms exist, such as ID3, C4.5, C5.0, and CART. CART, being the most recent and enhanced version, is employed in our model.

(a) **Gini impurity**: Used by the CART algorithm for classification trees, Gini impurity measures how often a randomly chosen element from the set would be incorrectly labelled if it were tagged at random in accordance with the subset's label distribution.
(b) **Information gain**: Utilized by the ID3, C4.5, and C5.0 tree generation algorithms, information gain is based on the concept of entropy and information content from information theory. It is used to determine which feature to split on at each step while constructing the tree.

Input: an attribute-valued dataset D
1: Tree = {}
2: **if** D is "pure" OR other stopping criteria met **then**
3: terminate
4: **end if**
5: **for all** attribute $a \in D$ **do**
6: Compute information-theoretic criteria if we split on a
7: **end for**
8: a_{best} = Best attribute according to above computed criteria
9: Tree = Create a decision node that tests a_{best} in the root
10: D_v = Induced sub-datasets from D based on a_{best}
11: **for all** D_v **do**
12: $Tree_v = C4.5(D_v)$
13: Attach $Tree_v$ to the corresponding branch of Tree
14: **end for**
15: **return** Tree

```
Decision Tree
Accuracy
0.9512195121951219
39
Confusion matrix
[[1 0 0 ... 0 0 0]
 [0 1 0 ... 0 0 0]
 [0 0 1 ... 0 0 0]
 ...
 [0 0 0 ... 1 0 0]
 [0 0 0 ... 0 1 0]
 [0 0 0 ... 0 0 1]]
['chest_pain', 'cramps', 'fast_heart_rate', 'belly_pain', 'back_pain']
[1, 1, 1, 1, 1]
```

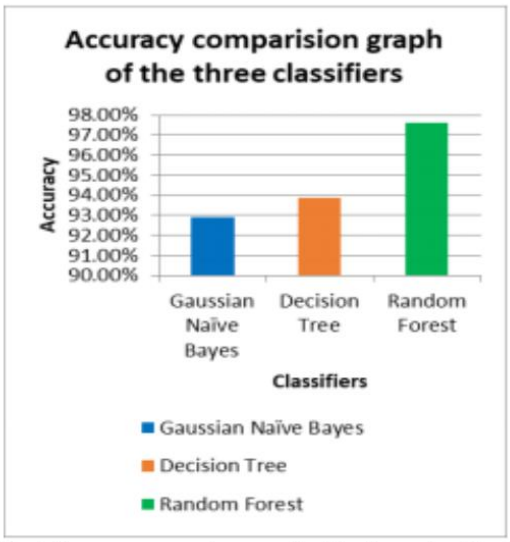
7. EVALUATING THE MODEL & RESULTS

The results produced by our model are summarised in the table below:

Algorithm	Accuracy before preprocessing	Accuracy after preprocessing
Gaussian Naïve Bayes	88.08%	92.9%
Decision Tree	90.12%	93.85%
Random Forest	95.28%	97.64%

[Table 3: Accuracy comparison table of the three algorithms]

It can be noted that the pre-processing technique, discretisation, has enhanced the performance of all three algorithms. While Naïve Bayes experienced a notable increase in accuracy due to discretisation, Random Forest achieved the highest accuracy for our dataset. The bar chart below illustrates that the Random Forest classifier outperforms the other two classifiers, making it the most suitable for our dataset:



[Fig 6: Accuracy comparison graph of the three classifiers]

8. CONCLUSIONS

The primary objective of this disease prediction system is to forecast diseases based on user symptoms. Users input their symptoms, and the system generates disease predictions with an average accuracy probability of 100%. The implementation of the Disease Predictor utilised the Grails framework, offering a user-friendly interface and ease of use.

Being a web-based application, users can access the system from anywhere and at any time, enhancing its accessibility and convenience. Overall, the accuracy of disease risk prediction relies on the diversity of features in the hospital data.

This systematic review aims to assess the performance, limitations, and prospects of software in healthcare, particularly focusing on Disease Predictability Software. The insights gained from this review can guide future developers in creating more effective and personalised patient care solutions. The program predicts patient diseases based on user input symptoms.

The system employs various algorithms such as Decision Tree, Random Forest, and Naïve Bayes to predict diseases. Machine learning algorithms process the data stored in the database, resulting in an accuracy rate of 98.3%. These machine learning techniques are specifically tailored to predict outbreaks successfully.

REFERENCES

[1] Pingale, K., Surwase, S., Kulkarni, V., Sarage, S., & Karve, A. (2019). Disease Prediction using Machine Learning.

[2] Aiyesha Sadia, Differential Diagnosis of Tuberculosis and Pneumonia using Machine Learning (2019)

[3] S. Patel and H. Patel, “Survey of data mining techniques used in healthcare domain,” Int. J. of Inform. Sci. and Tech., Vol. 6, pp. 53-60, March 2016.

[4] Balasubramanian, Satyabhama, and Balaji Subramanian. "Symptom based disease prediction in medical system by using K-means algorithm." International Journal of Advances in Computer Science and Technology 3.

[4] Dhenakaran, K. Rajalakshmi Dr SS. "Analysis of Data mining Prediction Techniques in Healthcare Management System." International Journal of Advanced Research in Computer Science and Software Engineering 5.4 (2015).

[5] DR. CK Gomathy, Article: A Semantic Quality of Web Service Information Retrieval Techniques Using Bin Rank A Cloud Monitoring Framework Perform in Web Services, International Journal of Scientific Research in Computer Science Engineering and Information Technology, (May-2018)

[6] D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, no G. Escobar, “Big data for health care: using analytics to identify and treat high-risk and high-risk patients, Health.

[7] K.R. Lakshmi, Y. Nagesh and Mr. Veera Krishna, “Comparison of performance of the three data mines ways to predict survival of kidney disease”, International Journal of Engineering Development & Technology, March 2014.