International Conference on Neural Nexus and Synergy:Innovation in Emerging Technologies
Vol.1.No.(2024). PP.275-286. SNIST

Smart Health Assistant: Integrating Machine Learning for Disease Prediction from Symptoms through Advanced Modelling

Koleti Ajay
B.Tech - IVth year Student,
CSE Department,
Sreenidhi Institute of Science and Technology(SNIST),
Hyderabad, India
ajaykoleti456@gmail.com

Aleti Dheeraj Kumar
B.Tech – IVth year Student,
CSE Department,
Sreenidhi Institute of Science and Technology(SNIST),
Hyderabad, India
20311a522@sreenidhi.edu.in

Kotha Rohith Reddy
B.Tech – IVth year Student ,
CSE Department,
Sreenidhi Institute of Science and Technology(SNIST),
Hyderabad, India
rohithreddykotha5424@gmail.com

Mr. D. Ram Babu
Assistant Professor,
CSE Department,
Sreenidhi Institute of Science and Technology(SNIST),
Hyderabad, India
rambabud@sreenidhi.edu.in

***Abstract*—**Using machine learning technology for disease prediction in real-world application in the field of medicine represents a revolutionary approach, giving new and previously impossible ways to save millions of lives. Machine learning models, relying on enormous databases and complicated algorithms, can find intricate patterns and connections that conventional diagnostic methods cannot find . In this work, we implement an advanced system of differential diagnostics based on machine learning platforms. We developed our classification models, such as a random forest gradient-boosting classification, and decision tree to achieve a high accuracy of the disease prediction. These models have high classification power due to their ability to work with considerable quantities and various types of data. First and foremost, the model has a number of clear and evident benefits; as a result, it allows individuals to receive timely and accurate predictions that will facilitate early recovery and treatment in cases when medical intervention is relevant. Second, with actionable data that indicates potential hazards for one's well-being, we can see that the system could significantly reduce the pressure on healthcare resources and minimize incremental rates of preventable diseases. Furthermore, the research emphasizes the outline and goal of predictive analytics in terms of introducing health efficiency and access to improvements. Therefore, with the help of predictive models that forecast the occurrence or development of disease, practitioners will become capable of better allocating resources, prioritizing individuals with a high likelihood of obtaining the disease, and adjusting treatment options to the needs of a specific patient.

Keywords—Gradient Boosting, Random Forest, Decision Tree, Machine Learning, Predictive Analytics, Gradient prediction.

I. INTRODUCTION

The capacity of machine learning to assess vast quantities of data and give insights that were once out of bounds has remained in high demand over the last few years . Furthermore, the

diagnostic, therapeutic, and administrative sectors of healthcare have recognized the role machine learning can play Patient data may be a bluff for precise, fast disease prediction. With patient data, machine learning offers an attractive opportunity for precise and effective disease prediction. Machine learning algorithms can improve diagnosis accuracy and patient results by identifying complex links in large-scale datasets. In order to forecast disease, the paper compares three machine learning techniques: Decision trees, random forests, and gradient boosting.

The objective is to assess how well illness prediction works with symptom information from medical records. The study uses rigorous analysis methods to train and assess these models. The performance of the model will be evaluated using important indicators such as classification reports, confusion matrix and accuracy.

This research has considered an area of disease prediction as one of the most crucial areas where machine learning can be significantly beneficial. The utilization of symptom data from large records and other self-reported sources creates an avenue to implementing and studying disease prediction. This work can be valuable in demonstrating the three major machine learning processes, and especially through systematic methods used in this research, this could be meaningful in comparing their performance and predictive accuracy for a disease.

II. LITERATURE SURVEY

Numerious studies have been carried out regarding disease prediction using various machine learning techniques which can be utilized by many medical facilities.

S. Grampurohit and C. Sagarnal[1] assessed machine learning models for diagnostic tests. They discovered that Support Vector Machine (SVM) outperformed Naive Bayes, Decision Trees, and KNearest Neighbor in terms of effectiveness for Parkinson's disease and renal ailment. Heart disease could be accurately predicted by using logistic regression (LR). CNN and Random Forest have demonstrated accuracy in predicting common ailments and breast cancer, respectively.

Aditi Gavhane[2] and her colleague proposed a machine learning model to predict cardiac disease. This system uses the multi-layer perceptron model. This model makes predictions on heart disease, based on some common symptoms like age, sex, pulse rate etc. The accuracy of the proposed system is 91%.

Gupta A, Kumar L, Jain R, and Nagrath P[3] proposed a heart prediction system utilizing the naive Bayes algorithm, achieving an accuracy of 97%. Based on symptoms such as breathing difficulties, back, neck, and chest pain, this algorithm makes predictions about cardiac illnesses. The system shows encouraging results in heart disease diagnosis by utilizing the ease of use and efficacy of the naive Bayes algorithm. This highlights the potential of machine learning to improve medical diagnosis and decision-making for better healthcare outcomes.

Sneha R,Nandini, Monisha S, Jahnavi C. “Disease Prediction System”.[4] 2021. Srms elibrary, 2021, Accessed 20 Oct 2021. The project by Iswarya et al. used classification algorithms to predict disease. Since Naive Bayes had accuracy of 97%, it has proven to be the suitable choice to use for disease prediction algorithm. Therefore, the use of various classification methods is critical to enhance overall disease prediction model accuracy and reliability.

Rudra[5] A and his colleague have suggested a system for multiple disease prediction . A remarkable novelty of this system is the anticipated appearance of consulting drug alcohol medication and medicine which isn’t there in this current model. The correctness of the said system is 85% .

Monika Gandhi and Shailendra Narayan Singh[6] proposed a framework designed to forecast heart ailments using data mining methodologies. Their investigation encompassed an examination of different data mining algorithms, such as Naive Bayes, Neural Network, and Decision Tree algorithms, implemented on medical data sets to enhance the accuracy of disease prognosis.

N. Kosarkar, P. Basuri, P. Karamore, P. Gawali, P. Badole, and P. Jumle[7] . The system developed using machine algorithms such as random forest, support vector machine, and logistic regression shows an accuracy of 82%. This system indicates that machine algorithms can also be used in predicting and making decisions about health outcomes.

Naveen Kumar, Naveenkumar S, Kirubhakaran R, Jeeva G, Shobana M, and Sangeetha Khas "Health Prediction System using Machine Learning Algorithms"[8]. They developed a health prediction system

using machine learning algorithms. With an accuracy rate of 94%, the system shows that machine learning techniques can be integrated into the health system to predict and diagnose medical conditions

Prediction of common diseases based on Dahiwade D, Patle, G, & Meshram,E[9] explored the prediction of common diseases utilizing patient symptoms, lifestyle habits, and diagnostic data through the application of KNearest Neighbour and Convolutional Neural Network. The findings demonstrated an accuracy of 84.5% for the CNN algorithm in forecasting common disease models, surpassing the accuracy of the KNN model.

Ambekar, S & Phalnikar, R[10] Accurate data analysis plays a vital part in the disease diagnosis and treatment, particularly at an early phase of patient care. Hence, using the Naïve Bayes and KNN algorithms, a Heart Disease's prediction model is developed over here, and later, it will be extended to predict disease risks from organized data.

III. BACKGROUND WORK

Indeed approaches to the integration of powerful technologies as modern machine learning have fundamentally changed approaches even to diagnosis and monitoring of patients . At the same time, these systems have several disadvantages that complicate their applicability and relevance in clinical practice.

Firstly, all approaches to the prediction of diseases and their criteria will be scientific data, and it is difficult for medical parameters to be evidence alone. The datasets themselves become outdated quickly, and the accrual and transmission of appropriate information will be difficult, especially in the regions with the lack of medical infrastructure.

Moreover, the use of such systems without the recommendation and appointment of a doctor will cause the user to select the wrong specialists, and non-existent, or nonconvenient, user interfaces without mobile support will make such services unavailable.

IV. METHODOLOGY

The project's goal is to develop an online platform that, using symptoms entered by users, employs machine learning to forecast likely symptoms. The website aims to empower individuals by offering user-friendly resources for preliminary health assessments, potentially resolving issues related to timely and cost-effective medical care access. The platform uses strong machine learning methods such as random forest, decision tree classification, and gradient boosting classification.

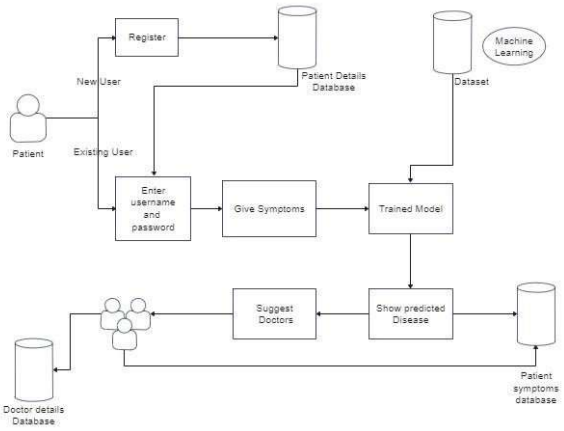

Figure 1: Architecture of Smart Health Assistant

In the given illustration, figure.1. Our system comprises three modules: Admin, User , and Doctor . A new user must register with the admin first. Upon a successful registration, the user will then have to enroll before signing in. A user only needs to sign up once. The illness prognosis system users are the physician, the patient, and the administrator. The system further verifies the identity of each and every user . System access is user-role-based. A patient can give symptoms, and the system will find the illness whereby the user will provide a probable diagnosis. The system also suggests a doctor once the illness has been predicted .The patient able to see a doctor online according to his convenience at home any free time.

A. Dataset Collection:

Data collection for disease prediction involves obtaining comprehensive datasets containing information about diseases and their associated symptoms. This dataset may consist of 133 columns and 4920 rows, providing detailed information about each diseasesymptom relationship.

itching	skin_rash	nodal_skin_eruptions	continuous_sneezing	shivering	chills	joint_pain	stomach_pain	acidity	ulcers_on_tongue	muscle_wasting
1	1	1	0	0	0	0	0	0	0	0
0	1	1	0	0	0	0	0	0	0	0
1	0	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0	0
1	1	1	0	0	0	0	0	0	0	0
0	1	1	0	0	0	0	0	0	0	0
1	0	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0	0
1	1	1	0	0	0	0	0	0	0	0
1	1	1	0	0	0	0	0	0	0	0
0	0	0	1	1	1	0	0	0	0	0
0	0	0	0	1	1	0	0	0	0	0
0	0	0	1	0	1	0	0	0	0	0
0	0	0	1	1	0	0	0	0	0	0
0	0	0	1	1	1	0	0	0	0	0
0	0	0	0	1	1	0	0	0	0	0
0	0	0	1	0	1	0	0	0	0	0
0	0	0	1	1	0	0	0	0	0	0
0	0	0	1	1	1	0	0	0	0	0
0	0	0	1	1	1	0	0	0	0	0
0	0	0	0	0	0	0	1	1	1	0
0	0	0	0	0	0	0	1	0	1	0
0	0	0	0	0	0	0	1	1	0	0
0	0	0	0	0	0	0	1	1	1	0

Figure 2: Data set 4920 records and 133 columns

B. Data Pre-Processing:

The purpose of data pre-processing is to organise and tidy up the data gathered. To maintain the consistency and uniformity of data, tasks include the removal of white spaces, punctuations and commas.

C. Classification:

Various machine learning classification methods are used to classify data and predict diseases based on specific symptoms. Gradient Boosting, Random Forest and Decision Tree are examples of common approaches.

1. **Random Forest:**
   This model, which falls under the category of supervised learning, employs marked data to enable the classification of unmarked data.. Unlike K-means cluster algorithms, which were discussed in previous articles as unsupervised learning models, Random Forest algorithms are versatile tools capable of solving regression and classification challenges and are the preferred choice among engineers.

Figure 2: Random Forest Algorithm

2. **Decision Tree:**

The Decision Tree Classifiers create a tree-like classification structure by recursively dividing data according to functional requirements. They pass from root to leaf node, represent the properties of each node, and describe the class label of each leaf node to find the best choice. Decision Tree Classifiers are unique in their adaptability and simplicity. Primarily, as they handle both numerical and categorical data, they are applicable in a myriad of classification problems in different applications domains. Moreover, their transparency makes it easy to visualize and interpret how the trees classify data and, by extension, discern the patterns within. Additionally, while the classifier might overfit with complex datasets, pruning ensures that the model generalizes better. On that note, Decision Tree Classifiers are fundamental in machine learning, given their intuitive and exhaustive approach in classifying data and learning useful patterns and relationships there in.

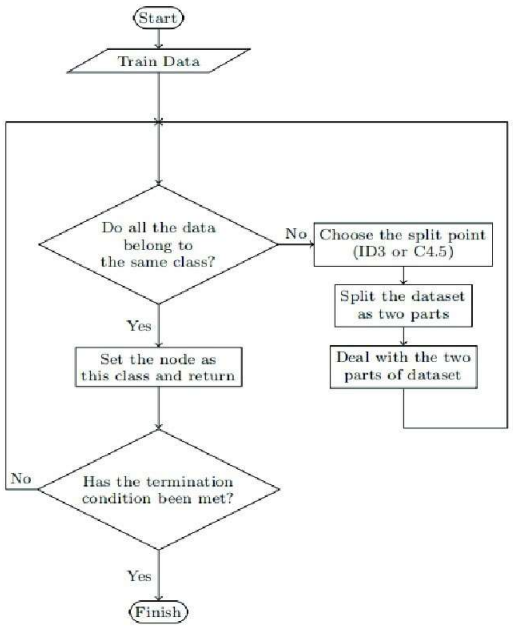

Figure 3: Decision Tree Algorithm

3. **Gradient Boosting:**

Gradient enhancement trains the decision tree team in sequence. Each tree corrects the error of the previous tree, focusing on the incorrect information of the previous tree. This collaborative approach makes it a powerful classification tool that achieves greater accuracy than a single tree. It can handle various data types and is less likely to be overused. Although powerful, it is complex, and requires more adjustment parameters than simple models.

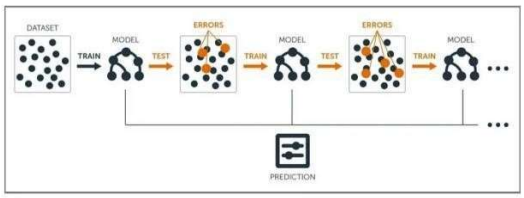

Figure 4: Gradient Boosting Algorithm

D. Model Training and Evaluation:

An essential step in the creation of a disease prediction system is the model training and evaluation phase. While fitting the training set to ML algorithms such as Decision Trees, Random Forests, and Gradient Boosting, pre processed data is used to let ML algorithms learn patterns and associations in the data. This is achieved by tuning internal parameters to minimize errors in predictions and improve predictive efficiency. Following fitting, each model is assessed using performance metrics to compare its effectiveness in omitted data, including accuracy, precision, recall, and F1-score. By using crossvalidation techniques, performance is determined while ensuring its generality and the latter model's power. Through repeating the crossvalidation and performance checking of the model, one can determine the best model for disease prediction.

E. Development of a web platform:

A user-friendly website was designed to allow users to input their symptoms and obtain the predictions of the disease. The frontend was done with help of HTML and CSS, allowing for building the interface, optimized for a userfriendly approach and pleasant appearance. The backend was developed with Django, a Python framework that allows for server-side logic to be handled, interaction with the model for disease prediction, and the training process itself. Additionally, the site is based on a PostgreSQL database management system. The website was developed with the opportunity for users to enter their symptoms through an input form. Therefore, after submission, the Django server on the backend processes the input, which then runs through the machine learning model. The predictions are then presented to the user, including relevant diagnosis conditions predicted for each disease. Ultimately, the platform supports users in predicting diseases based on symptoms in a simple and interactive way.

V. RESULTS AND DISCUSSION

After a detailed comparison of accuracy and crossvalidation scores in Gradient Boosting, Random Forest, and Decision Tree algorithms, Gradient Boosting was found the most efficient machine learning technique.

Provided that the achieved accuracy was 94.5%, the crossvalidation score was 97.6%, Gradient Boosting performed better than two other methods across both the accuracy of predictions and generalization levels. This result indicates the high prospects of using Gradient Boosting in disease prediction, which can become a revolutionary finding in medical theory and the spread of preventive medicine.

Model	Accuracy (before cross validation)	Accuracy (after cross validation)
Decision Tree	72.7%	76.1%
Random Forest	94.0%	95.2%
Gradient Boosting	94.5%	97.6%

Table 1 : Comparative study of performance

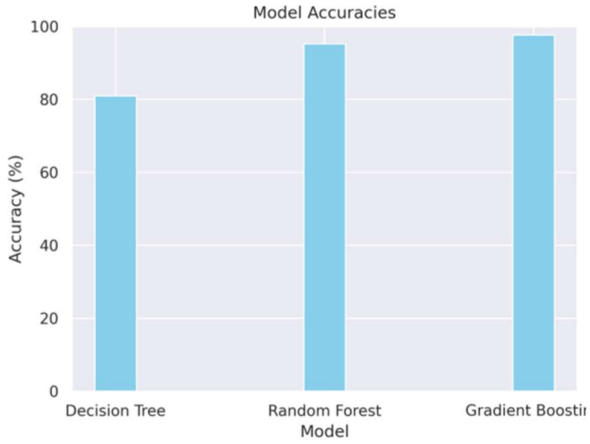

Figure 5: Accuracy of Models

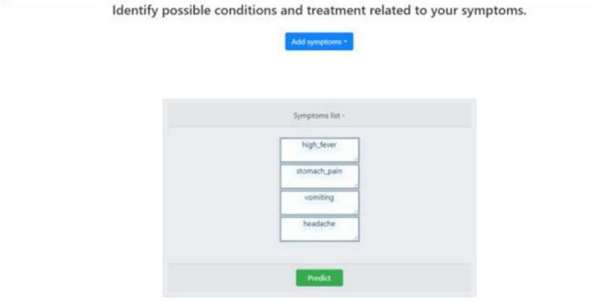

Figure 6.The user will add the symptoms

Figure 7: Output

VI. CONCLUSION

In conclusion, this project has presented the possibility of developing web-based platforms for disease prediction using machine learning technologies. Currently, the model's high performance level close to ideal, 97.6%, and the given condition of the platform provides the opportunity to launch. Its high level of accuracy proves the potential of this platform as a pre-health evaluation tool for consumers.

The findings and predictions presented by the platform in realtime can be beneficial to proactively intervene and make meaningful health-related decisions. However, it does not exclude the need for an on-line professional for doing a medical diagnosis. Therefore, the existing traditional medical intervention may be initiated during the consultation if the patient's indicators are alarming. Instead, the platform will complement the current healthcare by facilitating the transmission of knowledge that will lead to preventative medical practices.

VII. FUTURE SCOPE

Further tuning and refining of the machine learning models may improve their forecast accuracy and ability to generalize. This can include looking at cutting-edge techniques like hybrid models, deep learning, or ensemble learning to extract more intricate patterns from healthcare data.The developed prediction models can be integrated into existing healthcare systems and electronic health record (EHR) platforms to assist medical staff in promptly and accurately identifying patients. Public Health Initiatives,

preventive healthcare measures, such as lifestyle modifications or targeted screening programs.

REFERENCES

[1] Grampurohit, S. & Sagarnal, C., 2020. Disease Prediction using Machine Learning Algorithms. In: 2020 International Conference for Emerging
Technology (INCET), Jun 5-7, 2020, Belgaum, India

[2] Aditi Gavhane, Gouthami Kokkula, Isha Pandya and Kailas Devadkar,
"Prediction of Heart Disease Using Machine Learning" IEEE Xplore ISBN: 978-15386-0965-1, pp. 1275-1278, 2018.

[3] Gupta A., Kumar L., Jain R., Nagrath P. (2020) Heart Disease Prediction Using Classification (Naive Bayes). In: Singh P., Pawłowski W., Tanwar S., Kumar N., Rodrigues J., Obaidat M. (eds) Proceedings of First International
Conference on Computing, Communications, and Cyber-Security (IC4S 2019). Lecture Notes in Networks and Systems, vol 121. Springer, Singapore.

[4] Sneha R. Monisha S. Jahnavi C. and S. Nandini "Disease Prediction Based On Symptoms Using Classification Algorithm" Journal of Xi'an University of Architecture & Technology vol. 12 no. 4 2020.

[5] Rudra A. Godse Smita S. Gunjal Karan A. Jagtap Neha Mahamuni and Suchita Wankhade "Multiple Disease Prediction Using Different Machine Learning Algorithms Comparatively" International Journal of Advanced
Research in Computer and Communication Engineering vol. 8 no. 12 Dec 2019.

[6] Monika Gandhi and Shailendra Narayan Singh "Predictions in heart disease using techniques of data mining" 2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE) pp. 520-525 2015.

[7] N. Kosarkar, P. Basuri, P. Karamore, P. Gawali,P. Badole and P. Jumle, "Disease Prediction using Machine Learning," 2022 10th International Conference on Emerging Trends in Engineering and Technology - Signal and Information Processing (ICETET-SIP-22), Nagpur, India, 2022, pp. 1-4, doi:
10.1109/ICETET SIP2254415.2022.9791739.

[8] Naveenkumar S1 , Kirubhakaran R2 , Jeeva G3 , Shobana M4 , Sangeetha K5," Smart Health Prediction Using Machine Learning: A survey",3 march 2021.

[9] Dahiwade, D., Patle, G. & Meshram, E., 2019. Designing Disease Prediction Model Using Machine Learning Approach. In: 2019 3rd International
Conference on Computing Methodologies and Communication (ICCMC), March 2729, 2019, Erode, India.

[10]Ambekar, S.Phalnikar,R. , 2019, Disease Risk Prediction by Using Convolutional Neural Network,IN 2018 Fourth International Conference on Computing
Communication Control and Automation (ICCUBEA) 16-18 Aug. 2018, Pune, India.

[11] Prof. Shalu Saraswat, Shweta Gabhane, Alisha Pawar, Suhas Pingat, Shreyas Patil, "Implementation of Smart Health Prediction Using ML", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, PrintISSN : 2395-1990, Volume 10 Issue 3, pp.112-117, May-June 2023.

[12] S. Mahata, Y. B. Kapadiya, V. Kushwaha, V. Joshi, and Y. Farooqui,
"Disease Prediction and Treatment Recommendation Using Machine
Learning," International Journal for Research in Applied Science and Engineering
Technology, vol. 11, no. 3, pp. 1232–1237, Mar. 2023, doi: 10.22214/ijraset.2023.49641.