Smart Diagnosis: Symptom-Based Disease Prediction Using Machine Learning

Vijayakumar M[1], Navinbharathi M[2]
[1]Dept of Computer Science And Engineering
[2]Assistant Professor, Dept of Computer Science And Engineering
[1, 2]Madha Engineering College, Chennai-600069

Abstract- *The rapid developments in smart health care applications the usage of machine learning models, enhance the quality of disease prediction process, the outcome and the accuracy. The need for classifying the diseases from the massive data collected from health care infrastructure is difficult. The datasets are not similar and structured. Various challenges faced by the platform users to retrieve the analysis results within the time. the need for effective processing of healthcare data is demandable. The proposed system comprised of multiple machine learning algorithm comparisons to evaluate the performance of prediction quality as well as classify the electronic health care records (EHR) towards various disease category. Leveraging the predictive capability of the proposed model, the comparison towards performance, disease formulation is evaluated. The proposed system is implemented with authorized web application accessed by the health care professionals, to analyse the diseases. The quality of prediction is increased towards the quality of input dataset and preprocessing quality.*

Keywords: Health care, Internet of Things, Artificial Intelligence, Machine learning, Cognitive analysis.

I. INTRODUCTION

The development of the healthcare management system and the emergence of machine-learning algorithms utilised for prediction have increased the ability to provide effective diagnosis, prognosis, and treatment planning, making reliable solutions for maintaining healthcare records and providing timely support. Electronic health records (EHR) are large-scale medical datasets collected from various patients, allowing healthcare professionals to access them to identify patterns in the data. The primary objective of handling the healthcare dataset is to provide a timely solution to the various critical diseases, considering the patient's condition and historical Healthcare records. An electronic health record collects patient demographic information, genetic information, previous disorders, and timely updates of physiological data, which is recorded and maintained as a complete, structured database. Analysis of complex patterns and relationships between medical data collected from various patients and disease-related information, and the hidden correlations and genetic behaviours, using machine learning algorithms. The system can analyse patterns across various medical parameters to generate personalised insights for earlier intervention and analysis. The personalized health strategies are created as modules, offering the EHR, wi.th immediate suggestions to the patient metric analysis, so that, depending on the patient’s individual health care performance, provide a more accurate result. Various machine learning algorithms, regression algorithms, and ensemble learning approaches are used to develop robust predictive models for classifying specific diseases. Moreover, the integration of predictive analysis models into the electronic Healthcare recording system enables proactive interventions, regardless of a doctor's involvement. The computer can provide an artificial intelligence-enabled suggestion model to help with better decision-making. The power of Artificial Intelligence (AI) and Big Data Analytics enables predictive medicine, yielding highly accurate results that reduce manual assessment.

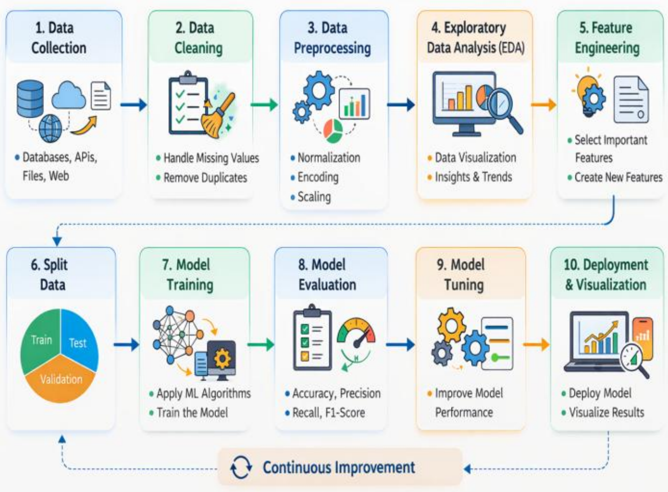


Fig 1. Steps involved in big data analysis

Fig 1. Shows the elements of big data analytics. The patient demographic data are preprocessed to handle missing values, normalize the features, and present the data scaling and provider quality data to improve feature extraction. The quality of this prediction model depends upon the quality of the feature extraction process. Using predictive analysis, the

system provides clinical decision-making and makes an efficient Health Care suggestion model using machine learning.

The primary motivation behind the disease prediction model persists with the advanced technology usage in the healthcare system, across various hospitals, in healthcare record maintenance systems, and increased reliability, provided by an automated framework with an automatic chat box. Leveraging the large data set collected from patients, the information needs to be associated with various diseases and analyzed using a machine learning model, enabling predictions and future interventions, and personalized treatment plans to be developed.

- The proposed system is focused on development of comparative analysis of machine learning models such as logistic regression, support vector machine, XGBoost model, long short term memory (LSTM) model etc.
- The input is nothing but real time collection of electronic health care records of various patients (EHR).
- The data is preprocessed, feature extracted and divided into training data, testing data. The data is fetched directly into the analysis model created with machine learning models.
- The performance of the system is evaluated through accuracy, precision, recall and F1Score.

The rest of the paper is formulated as existing scholarly articles learning in Section II, system tools required for implementation and dataset details in Section III, the design methodology and suggestion model-oriented discussions are made in Section IV etc.

II. BACKGROUND STUDY

N. Ghaniaviyanto Ramadhan et al. (2024) The author explored the chronic disease prediction system in which various physiological parameters are considered for analysis. The rapid developments of machine learning algorithm is more supportive for making the quality analysis of dataset. The major challenge exist within the big data analysis is the outliers, missing value replacement, presence of empty columns, imbalanced size of the data recordings. The quality of data plays an optimum role in influencing the accuracy of prediction systems. The chronic diseases exhibit resilient changes across the patient's physiological data and the metric parameters. The presented data make the analysis accurate and clear; the processing of the dataset before utilizing it in any application is highly recommended [6].

X. Yuan et al. (2022) The author explores a heart disease detection system and analysis model by analysing critical patterns among heart disease patients, developing early prevention, detection, and treatment procedures to protect them from life-threatening complications. The presented system used a boosted decision tree model supported by a Fuzzy Logic system to address data complexity and improve analysis, thereby avoiding overfitting. The presented system also utilised a bagging algorithm, in which multiple classifications are implemented. To accurately demonstrate the system, a stable multi-classification model is implemented [7].

F. Imrie et al. (2025) The author developed an exponential learning model to create various Diagnostic and prognostic approaches, helping build a deep decision-making model for a skin lesion dataset. The presented approach uses a convolutional neural network(CNN) architecture with vision Transformers to provide a multimodal approach for classifying the skin lesion problem. The presented system compared with existing state of art approaches. Primary challenge getting generated with the multi-modal dataset is explored and optimized for further enhancement of the work [8].

M. Irfan et al. (2023) The author explored the Alzheimer's disease (AD) detection framework using a machine learning algorithm. The presented system uses cognitive tests within simple machine learning approaches to provide multiple classifications of algorithms, achieving 93.92% with an artificial neural network model. The presented system detects the early stages of Alzheimer's disease using various Alzheimer's disease datasets and analyzes features, achieving an improvement of 12.12%. The primary study extracts the knowledge of machine learning models on handling large datasets[9].

Y. Shen et al. (2023) As various studies discuss the electronic Healthcare system, the primary role of clinical data is to improve the quality of disease prediction models. Using a simple deep learning algorithm, the presented system considered publicly available 6 datasets to analyse the presence of critical diseases and improve disease prediction performance. The mahcine learning model performance is evaluate through accuracy, precision,, recall and F1score. The difference between the actual data and the predicted data is analyzed [10].

III. SYSTEM DESIGN

The major problem with the disease prediction model is the large datasets, which are highly complex to process, and early predictions are difficult. Large high-resolution datasets consume more physical system space; hence, they require better handling of the feature extraction process. Accurate prediction and analysis of patient data help initiate diagnostic procedures early, preventing life-threatening problems. The current Medical System follows the conventional method of analysing diseases, in which most diseases recur with the same symptoms. The rapid development of adaptive learning algorithms enhances the quality of the disease analysis framework, enabling accurate analysis of large datasets of patient records. The challenges posed by most existing algorithms for logistic regression, the XGBoost algorithm, random forest, support vector machines, and the long short-term memory model are commonly used in machine learning to analyze patient health records. By understanding and mitigating the disease prediction model, personalised medical interventions are formulated to protect against the major problems arising from the diseases. Chronic diseases are evaluated based on the information provided, using historical data collected from various patients. The health care analysis is made, and the disease patterns that recur early are accurately predicted, so that the next patient experiences the same Diagnostic procedures and receives the same Emirates solution. The prediction models failed in the insertion and training process due to the lack of quality data.

IV. METHODOLOGY

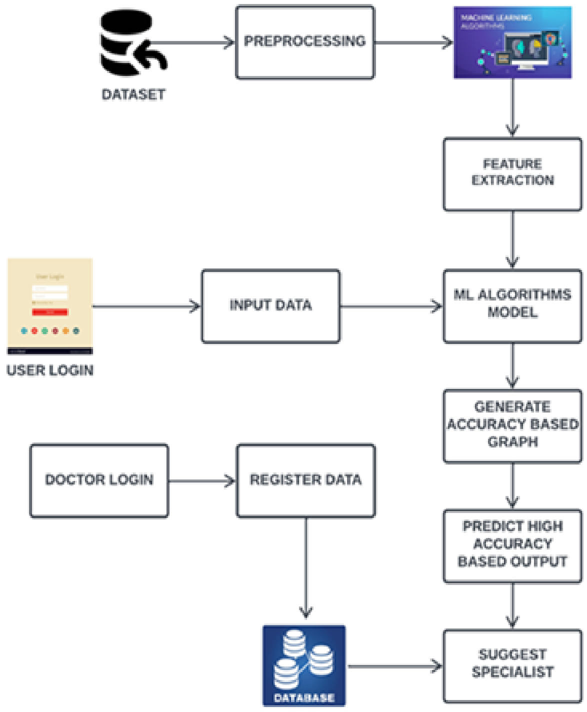


Fig. 2. System architecture

Fig 2. Shows the proposed system architecture consists of machine learning algorithm to leverage the prediction of human diseases based on the given dataset. A complete architecture is implemented by adding various machine learning algorithms to perform a comparative analysis, including the XGBoost regression algorithm, the random forest algorithm for stratification, the support vector machine (SVM), the logistic regression algorithm, and the LSTM neural network model, to accurately analyze the datasets. The capability of the proposed system is to make multiple decisions based on predictive analysis. The goal of the proposed approach is to determine which algorithm performs better for classification. Logistic regression is a fundamental classification method for the dataset and makes the feature matching approaches likely in a short time. Long short-term neural network employees perform temporal analysis of sequential data and identify patterns over time to make Complex decisions for the machine learning model, which provides the frequently occurring disease symptoms. Important health care records are essential to making the proposed system as successful as possible.

Preprocessing

Data processing is a crucial step in the decision-making of a disease prediction model, in which the key part ofthe dataset is to clean, combine the relevant information, make the analysis pattern, normalize the dataset, and extract the features in the form of symptoms, patient demographic data, and Medical Health records.

Feature Extraction

Feature extraction is a critical step in a disease prediction model, in which relevant information is extracted from the dataset to obtain deep-knowledge-driven features for classifying physiological data. The symptoms are accurately evaluated using temporal and time-series data. The dataset considers patterns that are recurrently extracted within the system.

Analysis model

The proposed machine learning algorithm evaluates the need to incorporate models such as the XGBoost, Random Forest, and logistic regression to classify the disease dataset. The ultimate goal of the system is to create an accurate model that can handle any dataset coming into the analysis application, classify them using the powerful learning and analysis framework, and use recurrent neural networks (RNNs).

XGBoost algorithm

XGBoost is an optimised distributed gradient boosting model that efficiently utilises parameters to achieve scalability. The extreme boost model comprises various machine learning algorithms in which the boosting process enhances the quality of the input dataset and mitigates random noise. The XGBoost model optimizes the dataset repeatedly to get patterns applicable for reducing the overfitting problems.

SVM algorithm

The support vector machine (SVM) is an unsupervised machine learning algorithm that utilises a feature space to map the data into a unique space. The number of features associated with the features that have unique information. are compared with various levels of disease prediction models, and finally, the decision-making processes is implemented.

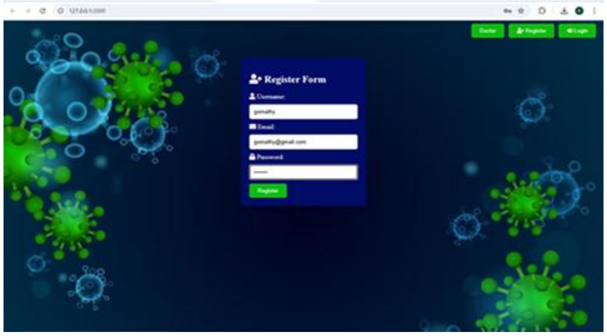
Logistic Regression

Logistic regression is a binary classification model in which the dataset is processed to produce class 0 and class 1, where the presence or absence of disease is classified as binary. The commonly used machine learning model in which the classification is held only as presence or absence.

LSTM model

The long short-term memory (LSTM) model is a recurrent neural network (RNN) architecture primarily designed to handle the complex structure of neural networks, with data complexity in mind. The neural network architecture considers the complex dataset and captures the patterns present in it; it scans for similar patterns and formulates the disease prediction. The LSTM model also impacts the processing time. The architecture's complexity increases the processing time required to understand the patterns.

V. RESULTS AND DISCUSSIONS



Register Form

Username

gomathy

Email

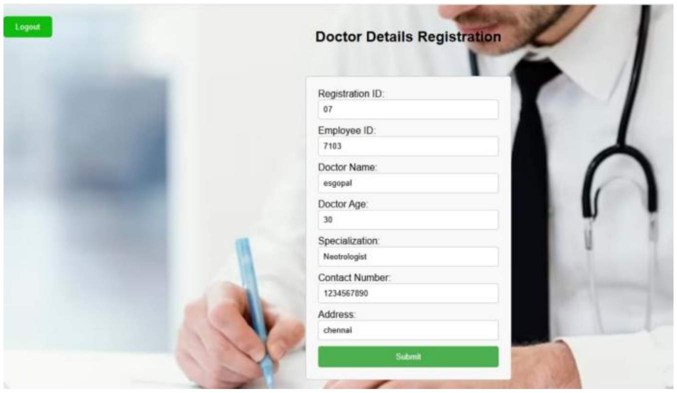
gomathy@gmail.com

Password

Register

Fig 3.Registeration forms

Fig 3. Shows the registration forms utilized for the propsoed disease analysis model. The users, patients opted to test the own records of historical meidcal records, get the suggetsion of doctors can utilize the page.



Doctor Details Registration

Registration ID: 07

Employee ID: 7103

Doctor Name: esgopal

Doctor Age: 30

Specialization: Neotrologist

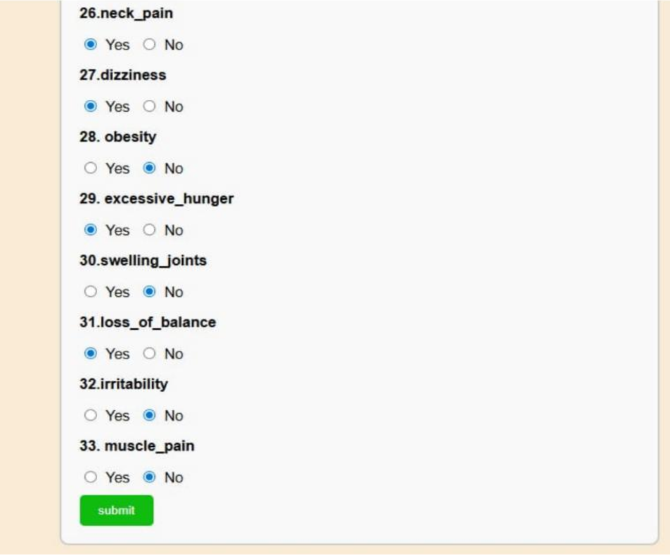
Contact Number: 1234567890

Address: chennai

Submit

Fig 4. Doctor’s registration

Fig 4. Shows the doctors registration window. It collects the doctor data, employee id, name, age, contact number, address are collected. The information is visbile to authorized users. The patient with user login authorized user can view the doctor details.



26.neck_pain

● Yes ○ No

27.dizziness

● Yes ○ No

28. obesity

○ Yes ● No

29. excessive_hunger

● Yes ○ No

30.swelling_joints

○ Yes ● No

31.loss_of_balance

● Yes ○ No

32.irritability

○ Yes ● No

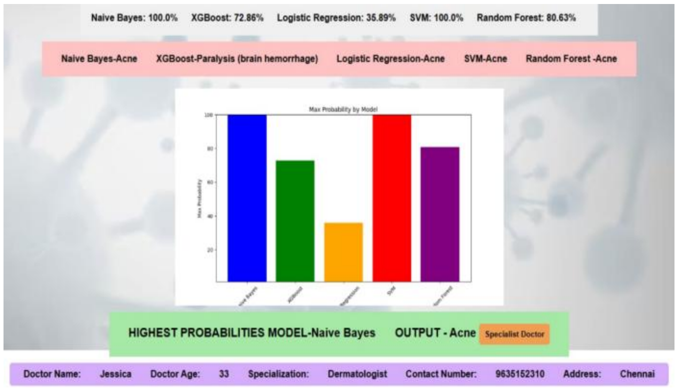
33. muscle_pain

○ Yes ● No

submit

Fig 5. Symptoms collection

Fig 5. Shows the data collection portal where various symptoms are collected from the patients. It collects various historical records about the patients helpful to integrate with analysis model.



Naive Bayes: 100.0% XGBoost: 72.86% Logistic Regression: 35.89% SVM: 100.0% Random Forest: 80.63%

Naive Bayes-Acne XGBoost-Paralysis (brain hemorrhage) Logistic Regression-Acne SVM-Acne Random Forest -Acne

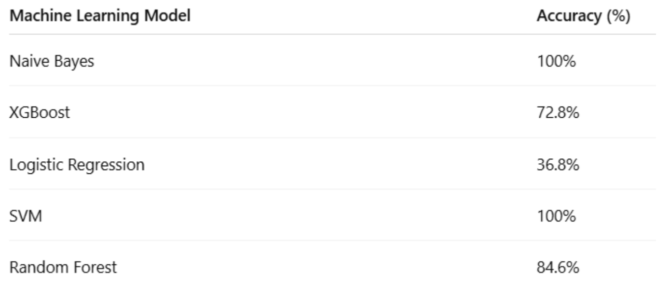
HIGHEST PROBABILITIES MODEL-Naive Bayes OUTPUT - Acne

Doctor Name: Jessica Doctor Age: 33 Specialization: Dermatologist Contact Number: 9635152310 Address: Chennai

Fig 6. Model probabilities and perfromance measure

Fig 6. Shows the model probabilities. The presented system achieved 98% accuracy towards the presented results. The result shows the detailed information regarding doctor authorized to view the report such as doctor name, age, specialization, contact number and address further the result also shows case the comparative output on given disease dataset catagorized as Acne, classified by the SVM model with 100% accuracy. The output is purely reply on given dataset.

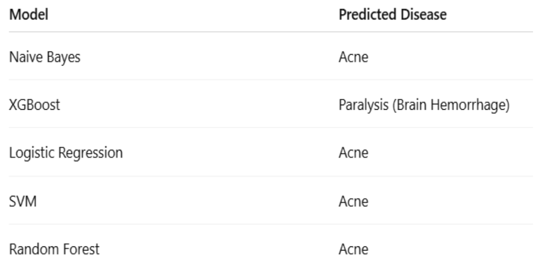
Table 1. Performance measure of proposed machine learning models



Machine Learning Model	Accuracy (%)
Naive Bayes	100%
XGBoost	72.8%
Logistic Regression	36.8%
SVM	100%
Random Forest	84.6%

Table 1. shows the performance measure of proposed machine learning models such as Naïve bayes model achieved 100% accuracy towards disease detection, XGBoost model achieved 72.8% accuracy, Logistic regression achieved only 36.8% accuracy, SVM model got 100% accuracy and Random forest scored 84.6% accuracy. Logistic regression model is lagging in performance for the particular input test dataset.

Table 2. prediction results



Model	Predicted Disease
Naive Bayes	Acne
XGBoost	Paralysis (Brain Hemorrhage)
Logistic Regression	Acne
SVM	Acne
Random Forest	Acne

Table 2. shows the prediction results of the proposed machine learning model classifying the input disease dataset as acne is the exact classification determined by the SVM model with 100% accuracy.

VI. CONCLUSION

After analyzing the dataset and employing various machine learning techniques, we have successfully developed predictive models for human disease diagnosis. By leveraging the wealth of data available, our models exhibit promising accuracy in forecasting various diseases. This advancement holds immense potential in revolutionizing healthcare, facilitating early detection and intervention. With these predictive capabilities, healthcare professionals can be equipped with valuable insights to better anticipate and manage diseases, ultimately leading to improved patient outcomes and quality of life.Furthermore, integrating these predictive models into clinical practice can significantly enhance the efficiency and effectiveness of healthcare delivery. By providing timely alerts and recommendations based on individual patient data, physicians can tailor their treatment plans and interventions more precisely. This personalized approach not only improves patient care but also optimizes resource allocation within healthcare systems, leading to better overall health outcomes and cost savings.In conclusion, the application of machine learning in disease prediction represents a significant stride forward in healthcare. By harnessing the power of data-driven insights, we can empower healthcare professionals with the tools they need to make informed decisions and provide personalized care to patients. This paradigm shift holds promise for a future where diseases can be anticipated and managed proactively, ushering in a new era of precision medicine and improved public health outcomes.

REFERENCES

[1] T. Exley, S. Moudy, R. M. Patterson, J. Kim and M. V. Albert, "Predicting UPDRS Motor Symptoms in Individuals With Parkinson's Disease From Force Plates Using Machine Learning," in IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 7, pp. 3486-3494, July 2022, doi: 10.1109/JBHI.2022.3157518

[2] R. Javed et al., "Enhancing Chronic Disease Prediction in IoMT-Enabled Healthcare 5.0 Using Deep Machine Learning: Alzheimer’s Disease as a Case Study," in IEEE Access, vol. 13, pp. 14252-14272, 2025, doi: 10.1109/ACCESS.2025.3525514

[3] S. T. Himi, N. T. Monalisa, M. Whaiduzzaman, A. Barros and M. S. Uddin, "MedAi: A Smartwatch-Based Application Framework for the Prediction of Common

Diseases Using Machine Learning," in IEEE Access, vol. 11, pp. 12342-12359, 2023, doi: 10.1109/ACCESS.2023.3236002.

[4] Alhumaidi, N. H., Dermawan, D., Kamaruzaman, H. F., & Alotaiq, N. (2025). The use of machine learning for analyzing real-world data in disease prediction and management: systematic review. JMIR Medical Informatics, 13(1), e68898.

[5] Zhong, J., & Wang, Y. (2025). Enhancing thyroid disease prediction using machine learning: A comparative study of ensemble models and class balancing techniques.

[6] N. Ghaniaviyanto Ramadhan, Adiwijaya, W. Maharani and A. Akbar Gozali, "Chronic Diseases Prediction Using Machine Learning With Data Preprocessing Handling: A Critical Review," in IEEE Access, vol. 12, pp. 80698-80730, 2024, doi: 10.1109/ACCESS.2024.3406748

[7] X. Yuan, J. Chen, K. Zhang, Y. Wu and T. Yang, "A Stable AI-Based Binary and Multiple Class Heart Disease Prediction Model for IoMT," in IEEE Transactions on Industrial Informatics, vol. 18, no. 3, pp. 2032-2040, March 2022, doi: 10.1109/TII.2021.3098306.

[8] F. Imrie, S. Denner, L. S. Brunschwig, K. Maier-Hein and M. van der Schaar, "Automated Ensemble Multimodal Machine Learning for Healthcare," in IEEE Journal of Biomedical and Health Informatics, vol. 29, no. 6, pp. 4213-4226, June 2025, doi: 10.1109/JBHI.2025.3530156

[9] M. Irfan, S. Shahrestani and M. Elkhodr, "Early Detection of Alzheimer's Disease Using Cognitive Features: A Voting-Based Ensemble Machine Learning Approach," in IEEE Engineering Management Review, vol. 51, no. 1, pp. 16-25, 1 Firstquarter,march 2023, doi: 10.1109/EMR.2022.3230820

[10] Y. Shen, J. Zhu, Z. Deng, W. Lu and H. Wang, "EnsDeepDP: An Ensemble Deep Learning Approach for Disease Prediction Through Metagenomics," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 20, no. 2, pp. 986-998, 1 March-April 2023, doi: 10.1109/TCBB.2022.3201295