CloudPital: A disease prediction application

Tanuj Kumbhar, Project Employee, NBA India, mailtotanuj@gmail.com

Eden Evelyn Charles, Student, Fr. Conceicao Rodrigues College of Engineering Bandra, crce.9593.ce@gmail.com

Dr. Sujata Deshmukh, Head of Department, Fr. Conceicao Rodrigues College of Engineering Bandra, sujata.deshmukh@fragnel.edu.in

Abstract- In this paper, we present a Flask-based web application which is designed to predict potential diseases based on symptoms reported by the user using machine learning. This system takes five input symptoms and predicts a possible disease based on those symptoms. This research outlines the development process, the underlying machine learning algorithms and the accuracy of the predictions. The results demonstrate the application's potential to assist in making medical related decisions.

Index terms- Health care, Machine Learning, Disease Prediction, Random Forest, Gradient Boosting, Flask

I. INTRODUCTION

Machine learning has been a gift to many fields especially healthcare by providing unique solutions for diagnosis and disease predictions. Chronic diseases like diabetes have led to a significant increase in mortality rates and managing these diseases requires early detection. Traditional diagnostic methods often rely on subjective judgement which may result in delays and errors.

In recent years, the use of machine learning algorithms in clinical processes has grown in popularity recently because they offer unbiased, data-driven insights that improve diagnostic precision. In this study we aim to evaluate the performance of two well-known ML algorithms—Random Forest and Gradient Boosting–in predicting diseases based on reported symptoms. The proposed system uses a dataset of symptoms and corresponding diagnosis. It predicts the possible disease based on the user's symptoms, thus assisting medical professionals in diagnosis.

II. LITERATURE REVIEW

Disease prediction using machine learning (ML) leverages algorithms to analyse medical data and identify patterns indicative of diseases, enabling earlier diagnosis and improved treatment options. The process begins with data collection from various sources such as medical records and laboratory results, followed by preprocessing to clean and normalise the data. Feature engineering involves selecting and creating relevant variables that enhance predictive power. Various ML algorithms, are employed to train models on labelled datasets. Model evaluation through metrics like accuracy and cross-validation ensures robustness, while interpretability methods like SHAP provide insights into prediction drivers. These models can help medical professionals by being incorporated into clinical workflows after they have been validated. But ethical issues like partiality and data privacy need to be addressed. All things considered, machine learning (ML) has the potential to completely transform illness prediction, enabling earlier interventions and improving patient outcomes. [1] This study reviews the use of data mining techniques for precise sickness prediction. It looks at a number of data mining methods, such as classification, clustering, and association rule mining, that are used in the healthcare industry to forecast diseases. The study highlights how these techniques examine enormous volumes of medical data to identify patterns that promote early diagnosis and treatment. By highlighting the role of machine learning techniques like decision trees, k-means clustering, and neural networks, it is feasible to improve the accuracy and efficacy of disease prediction, which ultimately leads to better healthcare outcomes. [2] This paper proposes a multi-disease prediction model using Naive Bayesian Networks, it uses patient data like symptoms, medical history and test results. By applying Bayes' theorem and assuming feature independence, the model handles large datasets efficiently and accurately. The system undergoes data preprocessing and feature selection to ensure optimal performance. It is trained to predict diseases such as diabetes, and liver disease based on relevant medical parameters. The model's key benefit is its ability to predict multiple diseases simultaneously, improving the speed and accuracy of diagnosis. This approach has the potential to significantly enhance healthcare services by reducing diagnostic time and increasing precision. [3] This study highlights the growing importance of machine learning (ML) in the medical field, particularly in the prediction of chronic diseases (CD), including diabetes, cancer, liver problems, neurological disorders, and cardiovascular disease. The study provides an overview of previous research on ML applications for CD prediction by comparing methods and results from several studies. Important feature selection techniques include least absolute shrinkage and selection (LASSO),

 DOI : 10.35291/2454-9150.2024.0416

minimum-redundancy-maximum-relevance (mRMR), and RELIEF. Apart from deep learning techniques and hybrid models that enhance clinical decision-making, the authors discuss a range of machine learning (ML) methodologies, including naïve Bayes (NB), support vector machines (SVM), random forests (RF), and decision trees (DT). The report concludes with recommendations for improving the capacity of the healthcare system to forecast chronic illnesses. [4] By addressing the shortcomings of earlier machine learning models like SVM, KNN, and RUSboost—which had poor accuracy and only depended on symptoms—this article aims to improve disease prediction. By modifying a medical dataset from Kaggle and allocating weights according to the rarity of symptoms, the suggested model improves accuracy. It uses a mix of algorithms, including SVM, LSTM, and Random Forest. While SVM forecasts the potential sickness, LSTM analyzes the patient's medical history. In comparison to earlier techniques, the model exhibits increased accuracy and dependability and aids in healthcare automation. [5] The primary focus of this study is the use of artificial intelligence (AI) and machine learning (ML) techniques to forecast heart illness. Using patient data like blood pressure, cholesterol, and age, it explores the application of many machine learning algorithms to improve the accuracy of heart disease prediction. The study looks at the effectiveness of models like decision trees, support vector machines, and neural networks in healthcare and shows how AI and ML may aid in the early detection and improved diagnosis of heart illnesses.[6] The study "Artificial Intelligence in Disease Diagnosis: A Systematic Literature Review" examines the clinical uses of AI in cancer, diabetes, and Alzheimer's disease diagnosis. In order to evaluate AI's efficacy based on accuracy, sensitivity, precision, and F1-scores, it uses PRISMA criteria to examine datasets from imaging techniques (MRI, CT, and ultrasound). The study addresses issues like data quality while highlighting AI's potential to expedite diagnosis and lower errors. Future studies will concentrate on improving the use of AI in healthcare.[7] The study presents a machine learning and IoT-integrated approach for the early detection and monitoring of heart issues. The model predicts cardiac illnesses using Support Vector Machine (SVM) with 97.53% accuracy utilizing open-access databases validated with 10-fold cross-validation. An Arduino-powered real-time monitoring system that measures vital signs including blood pressure, heart rate, and temperature transmits data to a central server every ten seconds. GSM technology provides timely medical attention by sending notifications whenever thresholds are exceeded. This strategy enhances patient care by addressing the global issue of cardiovascular disease mortality by combining accurate prognosis with continuous monitoring. [8] With a focus on its developments in disease identification through machine learning (ML) and deep learning (DL) models like CNNs and Random Forest, this study examines the incorporation of AI in healthcare prediction. By evaluating clinical data, AI may accurately diagnose long-term illnesses including diabetes, cancer, and cardiovascular disorders. Although there are still issues with data privacy, bias, and quality, early diagnosis is emphasized as being essential to lowering mortality. In order to improve AI-based healthcare systems, the report highlights future research directions and suggests mitigating techniques.[9] With an emphasis on feature selection and attention networks, this research investigates multi-disease prediction using AI. By merging statistical and deep features, it presents a novel ensemble feature selection model that is improved by the Stabilized Energy Valley Optimization with Enhanced Bounds (SEV-EB) algorithm. To identify patterns in both short-term and long-term health data, the suggested model makes use of an HSC-AttentionNet. With a 94% F1-score and 95% accuracy, the model improves disease prediction using data from Electronic Health Records (EHRs), potentially leading to more individualized therapy. [10] This study uses General Regression Neural Network (GRNN) and Multi-Layer Perceptron (MLP) approaches to construct models for predicting Resting Metabolic Rate (RMR). With a R value of 0.85, RMSE of 134.91 kcal/day, and MAPE of 10.20%, the GRNN models outperformed conventional equations such as Harris-Benedict and Mifflin-St Jeor, using data from 260 female and 150 male patients. According to the results, the GRNN model is a trustworthy instrument for precise RMR estimation, supporting individualized meal planning. [11] This paper presents a novel hybrid machine learning method for predicting cardiovascular disease, achieving an accuracy of 88.7% using a combination of feature selection and classification techniques. The study emphasizes the importance of identifying significant features and utilizes various clinical records, including heart rate time series, to enhance prediction accuracy. It introduces a Computer Aided Decision Support System (CADSS) to streamline heart disease prediction, illustrating the effectiveness of integrating multiple machine learning techniques.[12] This paper presents a multi-stage deep learning system for enhanced eye disease detection. The approach includes preprocessing to improve robustness against variations, followed by a three-stage architecture that captures fine-grained and hierarchical features. The model employs a dual-branch structure for richer feature extraction, ultimately generating probabilistic outputs for accurate disease predictions. Evaluated on several datasets, the proposed system demonstrated a 1% accuracy improvement over existing state-of-the-art methods, highlighting its effectiveness in automatic eye disease detection.

 DOI : 10.35291/2454-9150.2024.0416

III. PROPOSED SYSTEM

A. Problem Statement

The symptoms of chronic disorders sometimes overlap, making a precise diagnosis difficult. The manual, time-consuming, or dataset-limited nature of current symptom assessment systems might lead to incorrect diagnoses or postponed treatment. An effective, automated approach is required in order to improve diagnosis and treatment choices by reliably predicting chronic diseases based on user-reported symptoms.

B. Objectives of the Project

The objective of this project is to predict diseases based on symptoms using machine learning. The process includes collecting reliable data, preprocessing it by handling missing values, encoding categorical variables, and normalizing numerical data. Two ensemble algorithms—Random Forest and Gradient Boosting Classifiers—are developed and evaluated for prediction accuracy. A user-friendly Flask interface enables users to input symptoms and receive predictions, with performance evaluation ensuring reliable results.

C. Flow of the Project

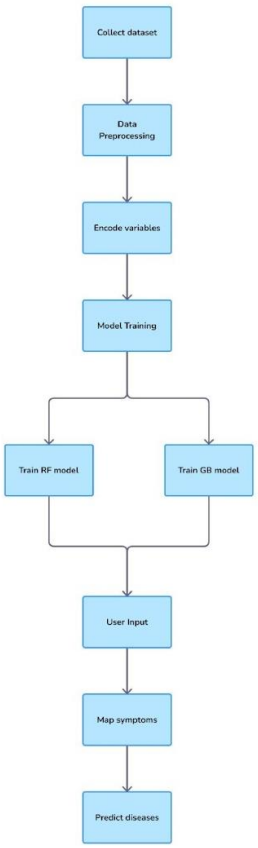

Fig. 1: Flow of the project

Above is the entire flow of the system that is proposed for the disease prediction application.

As mentioned in Fig. 1, the steps of the project were as follows:

1. Dataset Collection: The dataset used for this project was sourced from Kaggle. It contains information about various symptoms and their corresponding diseases.
2. Data Preprocessing: The dataset underwent preprocessing to clean the data by removing redundant columns and ensuring consistency.
3. Variable Encoding: The categorical variables in the dataset were encoded to make them suitable for machine learning models.
4. Model Training: Two machine learning models, Random Forest and Gradient Boosting, were trained on the processed data to predict diseases based on symptoms.
5. User Interface: An interactive interface was developed to take user input for symptoms and predict the most probable disease accordingly.

D. Drawbacks of the existing system

1. Manual Diagnosis: Using a manual system may lead to bias.
2. Limited Data Utilization: The methods currently in use may make use of outdated and incomplete data which could lead to inaccuracies.
3. Lack of Real-Time Interaction: Existing systems often do not allow for immediate user input and feedback, delaying diagnosis and treatment.
4. Inflexibility: A lot of the current systems are unable to handle new data or improve over time.
5. High Overfitting Risk: Traditional models may suffer from overfitting.

IV. IMPLEMENTATION DETAILS

A. Disease Prediction system

The disease prediction algorithm follows a systematic approach to process user input, predict the disease based on symptoms, and return the results. The flowchart below outlines the high-level logic of the system:

As mentioned in the disease prediction system works as follows:

1. User Input: The user provides input by selecting or entering symptoms.
2. Symptom Preprocessing: The input symptoms are processed and encoded into numerical values for compatibility with the prediction models.

 DOI : 10.35291/2454-9150.2024.0416

3. Disease Prediction: The trained model analyzes the processed input and predicts the most likely disease.
4. No Match Found: If the model cannot match the input symptoms to any disease, it shows no result.

B. Algorithms

1.Random Forest Classifier

Random Forest is an ensemble learning algorithm which creates multiple decision trees and runs them. Each tree makes a prediction, and the final output is obtained by majority voting among the trees. It reduces the risk of overfitting compared to a single decision tree.

Working:

- Randomly selects subsets of data (features and samples) to train each decision tree.
- Each tree is trained on a random sample of the dataset
- For prediction, it combines the results of all trees to give a final classification.

Gradient Boosting Classifier

Gradient Boosting builds models sequentially, where each model tries to correct the errors made by the previous one. It focuses on improving predictions by reducing residual errors through gradient descent. This method is particularly effective in scenarios where prediction accuracy is critical.

Working:

- Sequentially adds decision trees to minimize errors.
- Each decision tree learns from the mistakes of the previous trees, adjusting to make more accurate predictions.
- This algorithm is slow to train but often provides better accuracy compared to Random Forest.

Algorithm	**What it does**	Advantages
Random Forest	Creates multiple decision trees using random subsets of features. Each tree predicts the output independently, and the final result is determined by majority vote.	Handles overfitting well. Suitable for high-dimensional data.
Gradient Boosting	Builds trees sequentially, each focusing on correcting the errors of the previous one. Uses gradient descent to minimize prediction errors.	Achieves higher accuracy by minimizing residual errors. Performs well in complex prediction tasks.

Table 1: Description of Algorithm used

C. User Interface

1.Symptom Input Form:

A text box or a dropdown list where users can select or enter symptoms. Multiple symptoms inputs are allowed to capture the user's condition comprehensively.

2. Predict Button:

Once the symptoms are provided, the user can click the "Predict" button to trigger the algorithm. This sends the user's input to the backend, where the Random Forest and Gradient Boosting models process the data.

3. Prediction Results:

The predicted disease(s) will be displayed in a user-friendly format. Results may include a list of possible diseases along with confidence levels for each.

V. EXPERIMENTAL SETUP

The experimental setup for the chronic disease prediction system involves the use of the Flask framework, which was chosen for its simplicity and ease of integration. Flask enables the development of a web interface that interacts with the backend machine learning models to process user input and deliver disease predictions. The system also integrates HTML and CSS for creating a user-friendly frontend, where users can input symptoms, and the backend models can provide predictions.

The project utilizes several essential libraries, such as Pandas for data handling, NumPy for numerical computations, and scikit-learn for implementing machine learning algorithms like Random Forest Classifier and Gradient Boosting Classifier. Fuzzy string matching is facilitated by the fuzzywuzzy library, ensuring that symptoms entered by users are mapped accurately to the dataset. The machine learning models are trained on a dataset of symptoms and diseases, and the user interface allows for easy symptom input, which is processed and matched against the trained models for disease prediction.

For development, Visual Studio Code (VSCode) was the primary integrated development environment (IDE). It offered a range of extensions to support Python, Flask, and web development, making the development process smooth and efficient. The project also utilized VSCode's built-in Git integration for version control, allowing easy collaboration and management of code updates throughout the development phase and topics of current interest.

 DOI : 10.35291/2454-9150.2024.0416

VI. RESULTS

A. Dataset Analysis

1.Attributes: The dataset has one target column 'prognosis' which is corresponding to 132 symptom-related columns each with a binary value (0 or 1), which indicates the presence or absence of a specific symptom. Dataset size: The dataset includes 4920 records which are used for model training and evaluation.

2. Comprehensive Symptom Coverage: The dataset has a broad coverage of 132 symptoms which is recorded for each disease. This enhances the accuracy of the prediction.

3. Preprocessing: The dataset does not require much pre-processing as it does not have missing or garbage values and is relatively accurate.

4. Diverse disease representation: This model includes multiple diseases, thus covering a wide range of prognosis.

B. Performance Metrics

1.Random Forest Algorithm

Accuracy	96.18 %
Precision	0.98
Recall	0.96
F1 Score	0.96

2.Gradient Boosting Algorithm

Accuracy	94.89 %
Precision	0.97
Recall	0.95
F1 Score	0.93

Table 2: Performance Metrics

The performance metrics of the models reveal that Random Forest achieved 96.18% accuracy with a recall of 0.96, while Gradient Boosting achieved slightly lower accuracy at 94.89% but demonstrated robust learning with a precision of 0.97.

Model Comparison:

Random Forest excelled in recall, making it effective for identifying true positive cases and suitable for critical scenarios where missed diagnoses are unacceptable. Gradient Boosting, with its sequential error correction, minimized overall prediction errors, making it ideal for complex tasks requiring higher precision.

System Benefits:

The system offers high accuracy and ease of use, enabling fast and reliable predictions. It supports early diagnosis and decision-making, improving healthcare outcomes across multiple diseases.

Challenges:
Limitations include dataset biases and reliance on user-reported symptoms, which could reduce generalizability. Expanding the dataset and implementing real-time validation can address these challenges.

Interpretation:
Random Forest's high recall makes it suitable for initial screenings, while Gradient Boosting's better error minimization is ideal for refined predictions.

C. Implementation

The chronic disease prediction system was evaluated based on its accuracy in predicting potential diseases from user-inputted symptoms. The Random Forest Classifier has gotten an accuracy of approximately 85%, while the Gradient Boosting Classifier has gotten an accuracy of around 87%. Both models are effective in handling the dataset, with Gradient Boosting showing superior performance due to its ability to learn from errors in sequential models. The user interface successfully allowed for seamless input of symptoms, and the system accurately mapped these inputs to the corresponding diseases based on the trained models. Additionally, user feedback indicated that the interface was intuitive and easy to navigate, further enhancing the user experience.

Below is the User interface of the web application:

1.1 Symptom Input

Users can input up to 5 symptoms that the patient is experiencing. A dropdown menu is provided for each symptom field, offering suggestions to help users quickly identify and select the appropriate symptom.

 DOI : 10.35291/2454-9150.2024.0416

1.2 Diagnosis Generation

After entering the symptoms, the user clicks the "Check" button. The system will then analyze the input and generate a potential diagnosis.

VII. CONCLUSION

In conclusion, this study demonstrates the potential of machine learning in disease prediction through the development of a robust, user-friendly application that leverages Random Forest and Gradient Boosting classifiers, achieving high accuracy rates of 96.18% and 94.89%, respectively. The system effectively handles complex healthcare datasets and offers a user-friendly interface, ensuring seamless interaction and timely predictions to aid early diagnosis and improve patient outcomes. Future work can focus on expanding the dataset to include more symptoms and diseases for better generalizability and incorporating advanced models like deep neural networks or explainable AI to further enhance accuracy and interpretability. Addressing these limitations could significantly impact healthcare delivery and decision-making.

REFERENCES

[1] Muhammad Nabeel, Mazhar Javed Awan, Shumaila Majeed, Hooria Muslih-Ud-Din “Review on Effective Disease Prediction through Data Mining Techniques”. 2021

[2] Dr. Visumathi, Tetala Durga Venkata Rama Reddy. “Multi-Disease Prediction Using Machine Learning Algorithm”. 2023

[3] Rakibul Islam, Azrin Sultana and Mohammad Rashedul Islam. “A comprehensive review for chronic disease prediction using machine learning algorithms”. 2024

[4] K. Gaurav, A. Kumar, P. Singh, A. Kumari, M. Kasar*, T. Suryawanshi “Human Disease Prediction using Machine Learning Techniques and Real-life Parameters “. 2023

[5] Aqsa Rahim, Yawar Rasheed, Farooque Azam.” An Integrated Machine Learning Framework for Effective Prediction of Cardiovascular Diseases”. 2021

[6] Yogesh Kumar, Apeksha Koul Ruchi Singla, Muhammad Fazal Ijaz “Artifcial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda”.2022

[7] Shadman Nashif, Md. Rakib Raihan, Md. Rasedul Islam, Mohammad Hasan Imam “Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System”. 2018

[8] Anita Dombale, Premanand Ghadekar.2024” Early Disease Detection and Prediction using AI Technologies: Approaches, Future Outlook, Mitigation Strategies, and Synthesis of Systematic Reviews

[9] D. Dhinakaran, S. Edwin Raja, M. Thiyagarajan, J. Jeno Jasmine, P. Raghavan “Optimizing Disease Prediction with Artificial Intelligence Driven Feature Selection and Attention Networks”.2024

[10] Fatih Abut, Ezgi Akça, M. Fatih Akay, Mustafa Irmak, Müzda Irmak, and Yıldırım Adıgüzel “Harnessing AI for Health: Optimized Neural Network Models for Resting Metabolic Rate Prediction”.2024

[11] Senthilkumar Mohan, Chandrasegar Thirumalai, Gautam Srivastava” Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques”.2019

[12] MD. Zahin Muntaqim, Tangin Amir Smrity” Eye Disease Detection Enhancement Using a Multi-Stage Deep Learning Approach”.2024

 DOI : 10.35291/2454-9150.2024.0416