AI-Powered Disease Prediction and Personalized Health Recommendations

Team: Vital Vision
Members: Nikita(102317245),Aditya(102317258),Deepanshu Jindal(102317255)

Subject: Cognitive Computing
Thapar institute of engineering and technology,Patiala

March 26, 2025

1 Introduction

Early disease detection is essential for effective healthcare, but traditional diagnosis methods can be costly, time-consuming, and inaccessible to many. AI-driven solutions can bridge this gap by providing fast and reliable preliminary assessments.
This project presents a multi-disease prediction system that analyzes symptoms to identify possible diseases and recommends precautions, medications, and workout plans for better health management. Unlike single-disease models, which focus on specific illnesses, this system considers multiple conditions simultaneously, improving diagnostic accuracy and reducing misdiagnosis. The model predicts the top three possible diseases along with their probability percentages, helping users make informed decisions about their health.
By leveraging machine learning and medical data, it acts as a virtual health assistant, offering comprehensive and personalized guidance. While not a replacement for medical professionals, this tool enhances awareness, prevention, and self-care, making healthcare more accessible and efficient.
For implementation details and source code, refer to the GitHub repository: Source code

2 Dataset Overview

The dataset used for training and testing the multi-disease prediction model is sourced from Kaggle and consists of multiple CSV files. It includes **132 symptoms** mapped to **41 diseases** to ensure accurate and diverse predictions. The dataset comprises the following files:**Symptom-severity.csv**: Defines the severity levels of symptoms,**Training.csv**: Used as the training dataset for model development, **Symptoms_df.csv**: Used as the test dataset for evaluating model performance, **Description.csv, Diets.csv, Medications.csv, and Precautions_df.csv**: Contain disease-related descriptions, recommended diets, medications, and precautions and **Workout_df.csv**: Provides workout recommendations for different diseases.
To streamline recommendations, Description, Diets, Medications, and Precautions were combined into a single CSV file using the common column "Disease". This reference dataset helps generate personalized recommendations based on predicted diseases.
For implementation details and dataset access, refer to the Kaggle dataset: Kaggle_dataset.

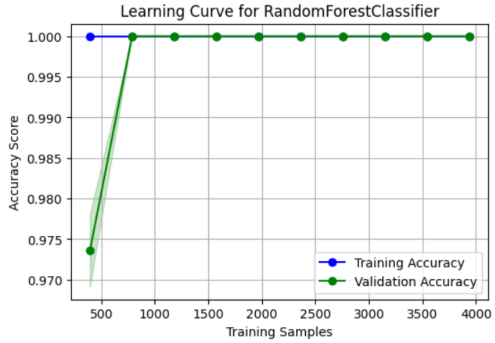
3 Technology Stack

The technology stack for this project is built around **Python** for data preprocessing, model training, and implementation. **Pandas** is used for data manipulation, while **NumPy** handles numerical computations. The model is trained and tested using three different machine learning approaches: Random Forest, XGBoost, and TensorFlow. After evaluation, the **TensorFlow** model was selected as the final model due to its superior performance. **Scikit-learn** is used for training and testing machine

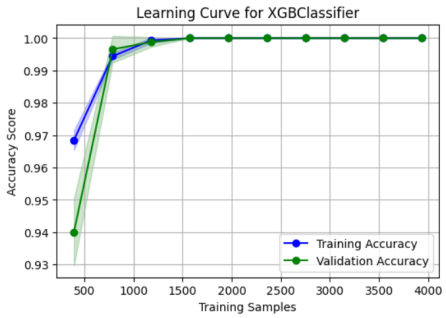
learning models, while **Matplotlib** is utilized for visualizing learning curves, accuracy comparisons, and other performance metrics. For deployment, **Pickle** is used to save and load the trained model, and **Gradio** provides an interactive user interface, enabling seamless deployment on Hugging Face. This technology stack ensures a robust and efficient multi-disease prediction system.

4 ML Model Implementation and Evaluation

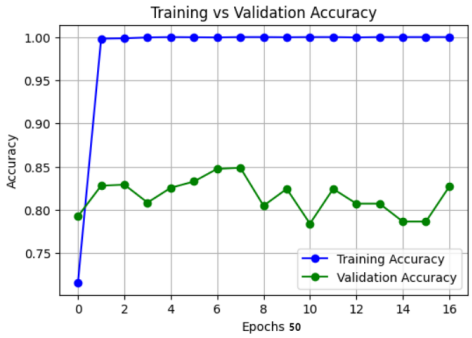
We experimented with multiple machine learning models to evaluate their performance on the given dataset. The models implemented include:**Random Forest Classifier,XGBoost Classifier,Deep Learning Model (Neural Network) with 50 and 100 Epochs**.
Each model was trained using a standardized dataset with appropriate feature selection, preprocessing, and hyperparameter tuning.
To assess the performance of our models, we utilized the following evaluation metrics:**Accuracy**: Measures the percentage of correctly classified instances. **Learning Curves**: Depicts how training and validation accuracy evolves with increasing training data. **Overfitting Analysis**: Examines the gap between training and validation accuracy.**Comparative Analysis of Models**: Evaluates models based on accuracy and generalization.



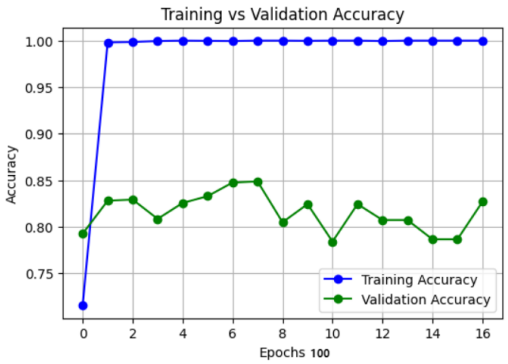
(a) Graph 1



(b) Graph 2



(c) Graph 3



(d) Graph 4

Figure 1: Evaluation Results of Different Training Epochs

4.1 Random Forest Classifier

The learning curve for the Random Forest model (Graph 1) shows that the training accuracy is consistently 100%, while validation accuracy stabilizes at nearly 99.9%, indicating potential overfitting.

4.2 XGB Classifier

The learning curve for XGBoost (Graph 2) exhibits a more gradual improvement, with validation accuracy reaching nearly 99.8%. It shows slightly better generalization than Random Forest.

4.3 Deep Learning Model

The deep learning model was trained for 50 epochs and 100 epochs (Graph 3 & 4). Initially, training accuracy rapidly increased to 100%, but validation accuracy fluctuated around 80% 85%, suggesting overfitting.
With 100 epochs, accuracy improved, achieving 84.76%, making it the best-performing model.

5 Results and Insights

A comparative accuracy analysis of all models is depicted in Figure 2. The results indicate: **Random Forest: 70.98%, XGBoost: 62.07%,Deep Learning (50 Epochs): 73.41%, Deep Learning (100 Epochs): 84.76%**. The deep learning model with 100 epochs performed the best, while XGBoost had the lowest accuracy.The model successfully predicted the top three most probable diseases along with their respective confidence scores, enhancing reliability.

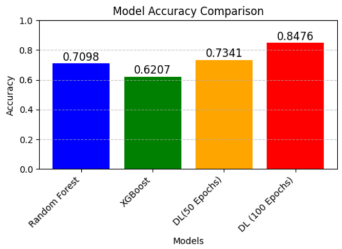


Figure 2: This graph depicts the accuracy of various models.

6 Challenges and Future Improvements

The challenges faced were that he deep learning model showed overfitting, requiring techniques like dropout and early stopping. Limited dataset diversity affected generalization, while imbalanced data led to biased predictions.Future improvements include expanding the dataset,Enhancing accuracy using transformer-based models and real-time symptom tracking via wearables can improve predictions.

7 Conclusion and Learning

This study highlights how model selection and training strategies impact accuracy. The deep learning model with 100 epochs performed best, showing the importance of training duration and dataset quality. Overfitting in Random Forest and XGBoost emphasized the need for proper validation techniques.

References

[1] https://www.kaggle.com/datasets/noorsaeed/medicine-recommendation-system-dataset

[2] https://www.youtube.com/watch?v=Mubj_fqiAv8&list=PLeo1K3hjS3uu7CxAacxVndI4bE_o3BDtO

[3]https://scikit-learn.org/stable/api/sklearn.html