DISEASE DIAGNOSIS FROM PATIENT MEDICAL RECORDS USING AI & ML(PYTHON)

G. Eepsitha Jahnavi[1], B. Harika[2],D. Syamala[3], D. Naresh[4], B. Venkat Sai Reddy[5]

[1]Assistant professor, [2,3,4,5]Students

[1,2,3,4,5] Department of Computer science and Engineering (Data Science),

[12,3,4,5]Avanthi institute of engineering of technology, Anakapalli,India.

{[1] eepsithajahnavig@gmail.com

[2] bylapudiharika27@gmail.com

[3] dalaboinasyamala76@gmail.com

[4] nareshdevara2003@gmail.com

[5] anki49383@gmail.com}@aiet.ac.in

1 ABSTRACT

Predictive Healthcare Analytics is an advanced medical decision-support system that utilizes artificial Intelligence (AI) and Machine Learning (ML) techniques to diagnose diseases from patient medical records. The proposed system is implemented using Python and Streamlit, integrating supervised learning algorithms such as Decision Tree, Random Forest, and Support Vector Machine (SVM).

The system processes structured medical datasets containing symptoms as input features and disease labels as output classes. After preprocessing and encoding, The dataset is split into training and testing sets. Multiple ML models are trained and evaluated using performance metrics such as Accuracy, precision ,Recall ,F1-score, and ROC Curve. The application provides an interactive web interface where users can select symptoms and obtain real-time disease predictions from multiple models. The system also compares model performances visually, helping identify the best-performing classifier. This approach enhances diagnostic accuracy, reduces manual effort, and supports early disease detection using intelligent analytics.

Keywords : Artificial Intelligence, Machine Learning, Disease Prediction, Healthcare Analytics, Decision Tree, Random Forest, Support Vector Machine (SVM), Medical Data Analysis, Predictive Healthcare, Streamlit.

2 INTRODUCTION

Healthcare systems generate large volumes of patient data daily, including symptoms, lab reports, and medical history. Analyzing this data manually can be time-consuming and prone to human error. With the advancement of Artificial Intelligence (AI) and Machine Learning (ML), automated systems can now assist doctors in diagnosing diseases based on patient symptoms and historical data.

Machine learning classification algorithms are widely used in medical diagnosis because they can identify patterns and correlations within large datasets. By training models on historical patient records, it becomes possible to predict diseases for new patients based on input symptoms.

This project focuses on building a multi-disease diagnostic system using Python. The system is capable of predicting diseases based on selected symptoms through a user-friendly Streamlit interface. The dataset is processed using preprocessing techniques such as label encoding and train-test splitting.

3 LITERATURE REVIEW

Artificial Intelligence (AI) and Machine Learning (ML) have played a significant role in improving disease diagnosis systems. Traditional methods relied on doctors' experience, patient history, and laboratory tests, which are time-consuming and prone to human error. Although Electronic Health Record (EHR) systems store patient data digitally, they lack predictive capabilities and intelligent analysis.

Machine learning algorithms such as Decision Tree, Random Forest, and Support Vector Machine (SVM) are widely used for disease prediction. Decision Tree is simple and easy to interpret, while Random Forest improves accuracy by combining multiple trees and reducing overfitting. SVM is effective for high-dimensional data and provides good classification performance. Studies show that ensemble methods like Random Forest generally perform better in medical datasets.

Modern healthcare systems use tools like Python, Scikit-learn, Pandas, and Streamlit for data processing, model building, and visualization. These technologies enable real-time disease prediction and performance evaluation. However, many systems still lack user-friendly interfaces and model comparison features. The proposed system addresses these issues by integrating multiple models with an interactive interface for efficient disease diagnosis.

4 PROBLEM STATEMENT

Healthcare systems generate a large amount of patient data, including symptoms, medical history, and clinical reports. However, traditional disease diagnosis methods rely heavily on manual analysis and the experience of doctors, making the process time-consuming and prone to human error. Existing systems such as Electronic Health Records (EHR) primarily focus on data storage and lack intelligent mechanisms for analyzing patient data and predicting diseases.

There is a need for an automated, accurate, and efficient system that can analyze patient symptoms and predict diseases using data-driven approaches. The challenge is to develop a system that not only provides reliable predictions but also compares multiple machine learning models and presents results in an easy-to-understand manner.

The proposed solution aims to address these limitations by designing an AI-based disease diagnosis system that uses machine learning algorithms to predict diseases from patient medical records, thereby improving diagnostic accuracy, reducing manual effort, and supporting healthcare decision-making.

5 SYSTEM ARCHITECTURE

The proposed system consists of three main components:

1. **Dataset Preparation** – A CSV file containing symptom indicators (0 or 1) and disease labels is used for model training.
2. **User Input Interface** – Interactive dropdown menus enable the user to select up to five symptoms. Input validation prevents duplicates and ensures at least one symptom is selected.
3. **Prediction Engine** – The selected symptoms are converted into a binary feature vector, which is processed by a machine learning model to predict the most probable disease.

6 METHODOLOGY

1. **Data Preprocessing** – Symptoms are encoded as binary values, and the disease column serves as the target variable.
2. **Input Validation** – The system ensures at least one symptom is selected and avoids duplicate entries.
3. **Prediction Process** – User-selected symptoms are converted into a numerical feature vector for disease prediction using a supervised learning model.
4.

7 TECHNOLOGY STACK

The proposed system is developed using Python as the primary programming language due to its simplicity and strong support for machine learning. Libraries such as Scikit-learn are used for implementing classification algorithms, while Pandas and NumPy are utilized for data preprocessing and numerical operations. Matplotlib and Seaborn are used for data visualization and performance analysis. The user interface is built using Streamlit, which enables real-time interaction and easy deployment of the application. Development and testing are carried out using tools like Visual Studio Code and Jupyter Notebook, making the system efficient and easy to manage.

The system is developed using the following technologies:

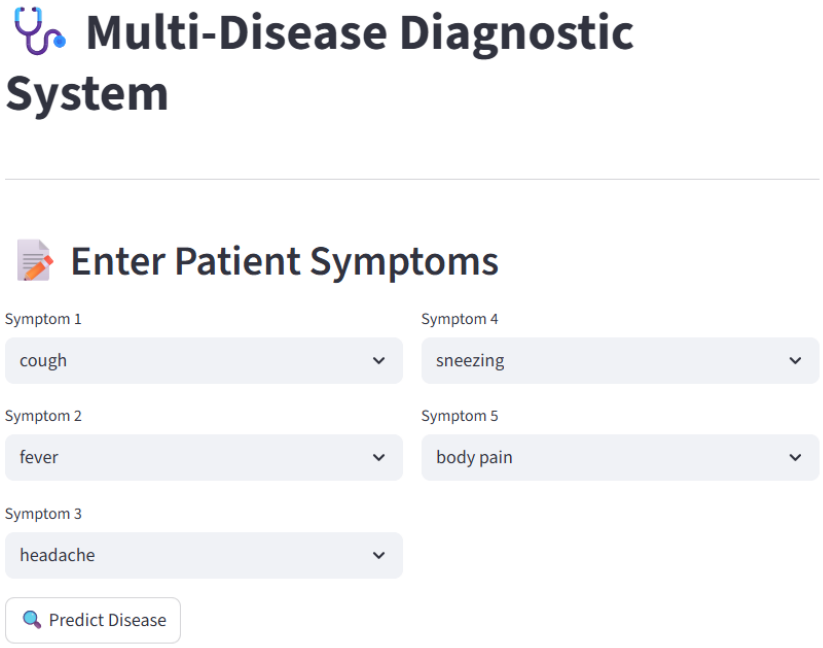
- **Programming Language**: Python
- **Machine Learning Library**: Scikit-learn
- **Data Processing**: Pandas, NumPy
- **Visualization**: Matplotlib, Seaborn
- **Web Framework**: Streamlit
- **Development Tools**: VS Code / Jupyter Notebook

These tools provide efficient development, analysis, and deployment.

8 INPUT DESIGN

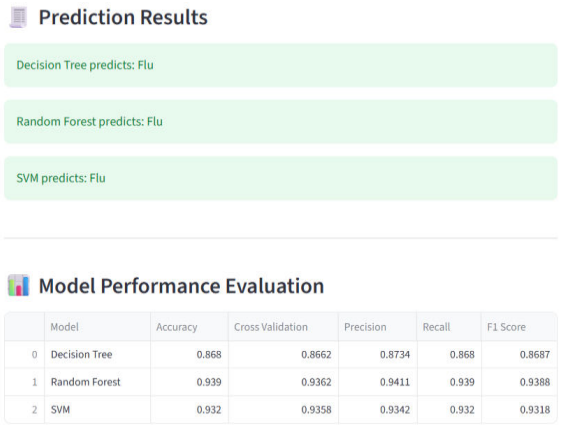
The input design of the proposed system focuses on accuracy, usability, and error prevention. Users can select up to five symptoms from interactive dropdown menus, which ensures that only valid entries are chosen and prevents manual input errors. The system incorporates input validation to guarantee that at least one symptom is selected and avoids duplicate symptom selections, maintaining data integrity. Once the user selects the symptoms, the system automatically encodes them into a binary feature vector (0s and 1s), aligning with the structured

format of the training dataset. This design ensures seamless integration between user input and the prediction model, enabling reliable and efficient disease prediction.



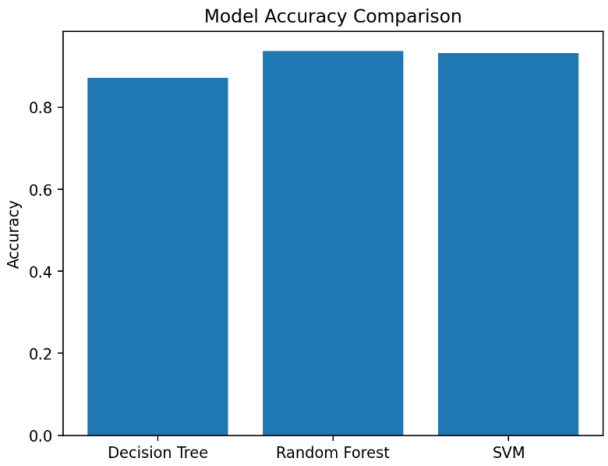
9 OUTPUT DESIGN

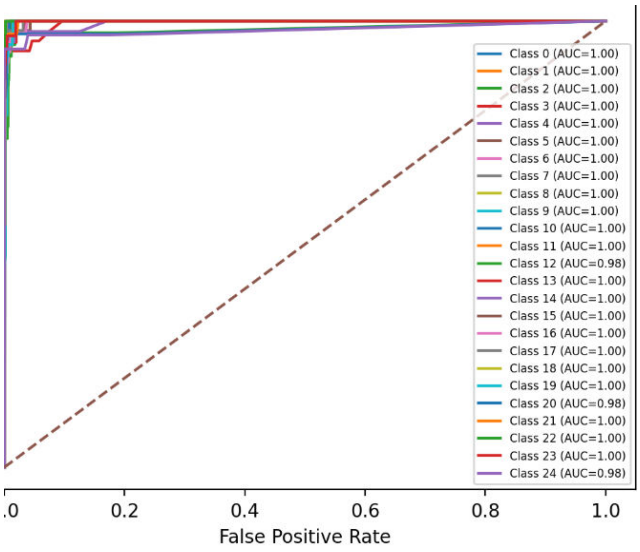
The output design of the system provides clear and actionable results for the user. After the symptoms are submitted, the machine learning model processes the binary feature vector and predicts the most probable disease. The predicted disease is displayed prominently on the interface, optionally accompanied by a probability score to indicate the confidence level of the prediction. The interface is designed to be user-friendly, ensuring that the results are easy to interpret, even for non-technical users. This design allows users to quickly understand the outcome and supports informed decision-making, enhancing the overall usability and effectiveness of the system.



Model Performance Evaluation

	Model	Accuracy	Cross Validation	Precision	Recall	F1 Score
0	Decision Tree	0.868	0.8662	0.8734	0.868	0.8687
1	Random Forest	0.939	0.9362	0.9411	0.939	0.9388
2	SVM	0.932	0.9358	0.9342	0.932	0.9318





10 ADVANTAGES

The proposed system offers several advantages, including faster disease prediction and reduced dependency on manual analysis. It minimizes human error by providing data-driven predictions and supports healthcare professionals in decision-making. The system is user-friendly and provides real-time results through an interactive interface. It also allows comparison of multiple machine learning models, enabling the selection of the most accurate model. Overall, the system improves efficiency, accuracy, and reliability in disease diagnosis.

11 FUTURE SCOPE

The future scope of the proposed system includes integrating it with real-time hospital databases to enable live data analysis. Advanced techniques such as deep learning models can be incorporated to improve prediction accuracy further. The system can also be extended into a mobile application for wider accessibility. Additional diseases and symptoms can be included to make the system more comprehensive. Furthermore, cloud-based deployment and integration with IoT healthcare devices can enhance scalability and real-time monitoring capabilities.

12 CONCLUSION

The proposed system successfully demonstrates the application of Artificial Intelligence and Machine Learning in disease diagnosis. By using classification algorithms such as Decision Tree, Random Forest, and SVM, the system predicts diseases based on patient symptoms with good accuracy.

The Streamlit interface makes the system interactive and easy to use, while performance metrics ensure reliability. The system reduces manual effort and provides quick, data-driven predictions. Although it cannot replace doctors, it serves as an effective decision-support tool in healthcare.

13 REFERENCES

1. Mitchell, T., *Machine Learning*, McGraw-Hill, 1997.
2. Goodfellow, I., Bengio, Y., & Courville, A., *Deep Learning*, MIT Press, 2016.
3. Raschka, S., *Python Machine Learning*, Packt Publishing, 2017.
4. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research (JMLR)*, 12, 2011, pp. 2825–2830.
5. McKinney, W., "Data Structures for Statistical Computing in Python," 2010.
6. Official Python Documentation, Available: https://www.python.org
7. Scikit-learn Documentation, Available: https://scikit-learn.org
8. Streamlit Documentation, Available: https://streamlit.io