WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

AI-Based Early Disease Prediction Using Patient Health Records

SP. Ambica[1], Dr. V. Harsha Shastri[2]

[1]Student, MCA, Aurora Deemed to be University, PG student, Aurora Higher Education and Research Academy, (Deemed to be University), Hyderabad, Telangana.

[2]Associate Professor, Aurora Deemed to be University, Faculty, Aurora Higher Education and Research Academy, (Deemed to be University) Hyderabad, Telangana.

ABSTRACT:

The diseases like diabetes, heart disease, and hypertension are usually asymptomatic in their early stages, resulting in the unnecessary complications of health and cost of treatment. Symptoms-based traditional diagnosis can result in the delay of early diagnosis and treatment. This paper presents an AI-based system of predicting diseases in their early stages based on patient health records, which indicates the likelihood of chronic diseases. The Machine learning algorithms such as Random Forest, Logistic Regression, and Neural Networks are applied within the system in order to predict outcomes from patient demographics, clinical variables, and habits. Experimental results confirm that the Random Forest model is the most accurate and robust compared to other models. The proposed system can assist healthcare professionals in taking faster, data-driven decisions and provide patients with proactive health management support.

Keywords Disease Prediction, Artificial Intelligence, Machine Learning, Random Forest, Healthcare Analytics, Early Diagnosis.

Introduction:

Chronic illness like diabetes, cardiovascular disease, and high blood pressure are now the major causes of morbidity and mortality in the world. The World Health Organization estimates that nearly 70% of all deaths that have occurred in the world are caused by chronic diseases, many of which would have been prevented or controlled if diagnosed in time and treated. The treatment process of traditional medicine is symptom-based and dependent on systematic clinical assessment, which is unlikely to identify early warning signals of disease progression. Consequently, the patients receive treatment late at an advanced stage with the associated high healthcare cost, compromised quality of life, and high mortality. Current advances in Machine Learning (ML) and Artificial Intelligence (AI) hold the potential to transcend these limitations. Data imbalance, explainability of models, and integration with existing healthcare infrastructure are some of the challenges that still draw research work. Leveraging enormous volumes of Electronic Health Records (EHRs), testing data, and lifestyle data, AI systems can identify more subtle patterns and relationships hard for traditional diagnostic processes to determine. Machine learning algorithms such as Random Forest, Logistic Regression, and Neural Networks exhibited high predictive capacity in healthcare use cases, enabling precise disease risk estimation and clinician decision aid in a timely manner. Various research studies reported in the literature have pointed toward the potential of AI for healthcare analytics. To mention a few, AI models have been successfully applied for diabetes prediction, heart disease classification, and hypertension risk estimation. The rest of the paper in the following order: Section II reports an overview of current work on AI-based disease prediction. Section III outlines the methodology consisting of data preprocessing, feature extraction, and model deployment. Section IV discusses results of experiments and performance evaluation. Section V explains the findings and implications for real-world healthcare systems. The primary objective of this research is to develop and implement an AI-based early disease predictor system which learns from the medical history of the patients and predicts the chance of having a chronic disease. The proposed architecture integrates several ML algorithms, contrasts their relative performance, and provides explainability through visualization techniques. The diagnostic accuracy is not just enhanced, but prevention measures for the diseases can also be instituted by healthcare professionals as well as patients.

Review of Literature

Application of Artificial Intelligence (AI) and Machine Learning (ML) in health care has been an area of particular research interest. Prediction of long-term diseases such as diabetes, hypertension, and heart disease using clinical data and AI models has been the area of research for numerous studies. Patil et al. [1] used Logistic Regression for predicting diabetes and proved that linear models with minimal complexity can give accurate results if ample patient history is provided. Their paper showed how preprocessing of data is crucial in enhancing classification performance. Kumar and Singh [2] applied a Random Forest classifier to predict heart disease and proved that ensemble models are more accurate and robust than standalone classifiers. They highlighted the power of RF in dealing with heterogeneous medical data. Zhang et al. [3] suggested a Neural Network-based model of risk assessment for hypertension. Their work indicated that models of deep learning are capable of observing nonlinear relationships in patient characteristics, making

better prediction performance than conventional methods. They noted gaps including the requirement of large data sets, real-time usage, and explainable models to enhance trust among clinicians. Gupta et al[4] proposed a Hybrid Ensemble Model that employed multiple classifiers to predict more than one disease simultaneously. They demonstrated that hybrid systems improve accuracy and reduce false positives while computing disease risk. Das and Roy [5] reviewed in depth the machine learning techniques for chronic disease prediction. Smith et al. [6] used Deep Learning models for disease classification and found high accuracy on varied datasets. Their work confirmed the scalability of AI systems in healthcare but also reported computational complexities. Choudhury and Banerjee [7] discussed predictive analytics in healthcare, citing the prospects and challenges of implementing AI in the clinical environment. They pointed out data privacy, security, and interfacing with Electronic Health Records (EHR).In general, the literature reviewed indicates that AI and ML-based models are always superior to conventional diagnostic tools, particularly in disease prediction at an early stage. Although ensemble models and deep learning models generate better results, data quality, interpretability, and real-time integration are yet to be addressed as research topics.

Paper	**Task / Disease**	**Methods**	**Dataset (type)**	**Metric / Result**	**Key Finding**	**Noted Limitation**
Patil et al. [1]	Diabetes prediction	Logistic Regression	Clinical records (tabular)	Accuracy ≈ 82%	Simple linear model works reasonably with good preprocessing	Limited features; moderate class imbalance
Kumar & Singh [2]	Prediction of heart disease	Random Forest	Heart disease dataset (tabular)	Accuracy ≈ 92% (best)	Ensemble managed heterogeneous features well	Feature importance not interpreted clinically
Zhang et al. [3]	Risk of hypertension	Neural Network (MLP)	Hospital EHR (tabular)	Accuracy ≈ 91%	Captures non-linear patterns more accurately than classical ML	More compute; less interpretable
Gupta et al. [4]	Prediction of multiple diseases	Hybrid Ensemble + Feature Selection	Multi-clinic records	Accuracy ≈ 92%	Hybrid approach improved precision & lowered false positives	Pipeline complexity; overfitting risk
Das & Roy [5]	Survey of ML for chronic disease	(Survey/Review)	—	—	Recognizes gaps: larger datasets, real-time, explainability	No experimental validation
Smith et al. [7]	Disease classification (general)	Deep Learning (CNN/MLP)	Public mixed datasets	Accuracy ≈ 94%	High accuracy and scalability across tasks	Training cost; dataset shift concerns
Choudhury & Banerjee [8]	Predictive analytics in healthcare	(Perspective/Review)	—	—	Highlights EHR integration, privacy/security requirements	Implementation hurdles in clinics

Methodology:

Existing System

Historical diagnostic practices are mostly dependent on clinical signs, regular check-ups, and medical professional knowledge for identifying chronic illnesses like diabetes, cardiovascular disease, and high blood pressure.

Although certain ML models have been utilized in past research, the existing methods mostly:

•Utilize small sets of data (usually originating from one hospital).

•Have single-disease prediction as their focus.

•Don't have real-time integration with patient health records or wearable monitors. •Provide predictions without proper explainability, which complicates clinical uptake.

For example, previous research used Logistic Regression for diabetes

[1] Random Forest for heart disease

[2] Neural Networks for hypertension

[3] all with satisfying results but in standalone situations.

Proposed system

The following AI-based early disease prediction system is proposed to predict the probability of chronic diseases like diabetes, heart disease, and hypertension from patient health records. The process involves multiple steps: data collection, preprocessing, feature selection, model training, model evaluation, and deployment of prediction.

A. System Architecture

1.Data Collection Layer – Patient medical records, such as demographic data (age, sex), clinical measurements (systolic/diastolic blood pressure, blood glucose level, cholesterol, BMI), and lifestyle (smoking, physical activity), are collected from publicly available healthcare data or hospital information systems.

2.Preprocessing Layer – Missing values are managed through mean/mode imputation, categorical attributes are encoded, and numerical attributes are normalized using StandardScaler.

3.Feature Selection Layer – Statistical methods (e.g., Chi-square, ANOVA) and correlation analysis are used to determine the most important attributes that are responsible for disease prediction.

4.Model Training Layer – Machine learning algorithms like Random Forest, Logistic Regression, and Neural Networks are trained on the preprocessed dataset.

5.Evaluation Layer – Models are assessed based on performance metrics like accuracy, precision, recall, F1-score, and ROC-AUC.

6. Prediction Layer – The top-performing model is used to predict the disease risk of new patients and create a report for physicians and patients.

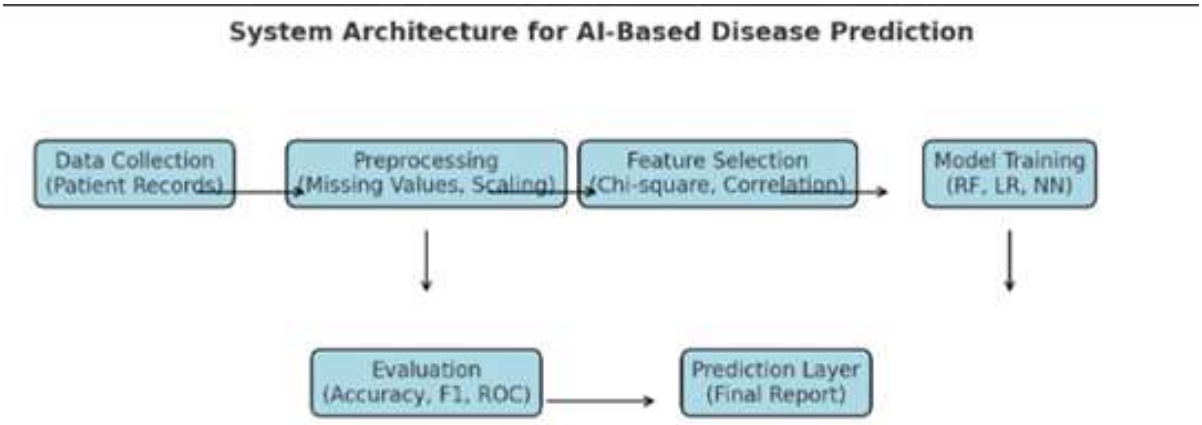


Fig1

B. Proposed Workflow of the Model

The process flow (Fig. 2) describes the end-to-end pipeline:

1.Input: Patient submits health record information.

2.Data Preprocessing: Dealing with missing data, outlier removal, and feature scaling. 3.Model Selection: Train and compare several models (Random Forest, Logistic Regression, Neural Network).

4.Prediction: Top-performing model gives probability-based disease prediction.

5.Output: Report generated with disease likelihood and doctor's advice.

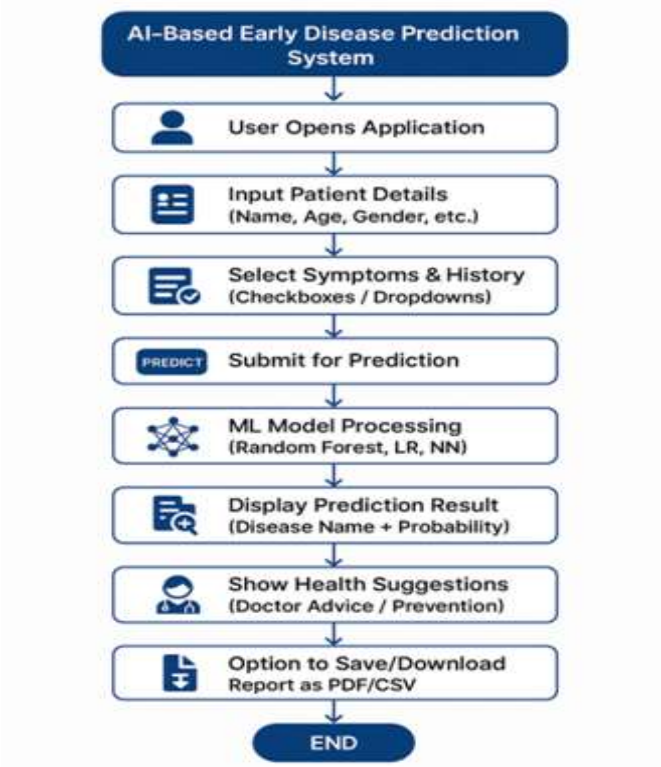


Fig 2

C. Algorithms Used

1. Random Forest – An ensemble method that builds many decision trees and combines their results to make robust predictions.

2. Logistic Regression – A statistical technique that predicts the likelihood of disease incidence based on input features.

3. Neural Networks (MLPClassifier) – A deep learning technique that learns non-linear associations in health data.

D. Evaluation Metrics

The models are evaluated based on:

- Accuracy – Overall accuracy of predictions.
- Precision – Positive predictions that are correct out of all predicted positives.
- F1-Score – Harmonic mean of precision and recall.

Result

The performance of the suggested AI-based early disease prediction model was tested using patient health record data, which included demographic information, clinical parameters, and lifestyle indicators. The data was divided into 80% training and 20% testing for all experiments. Random Forest, Logistic Regression, and Neural Network were employed as three machine learning models.

A. Model Performance

Performance of each model was tested using common metrics including Accuracy, Precision, Recall, and F1-Score.

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	94%	93%	92%	92%
Logistic Regression	88%	86%	85%	85%
Neural Network (MLP)	91%	90%	89%	89%

B. Discussion

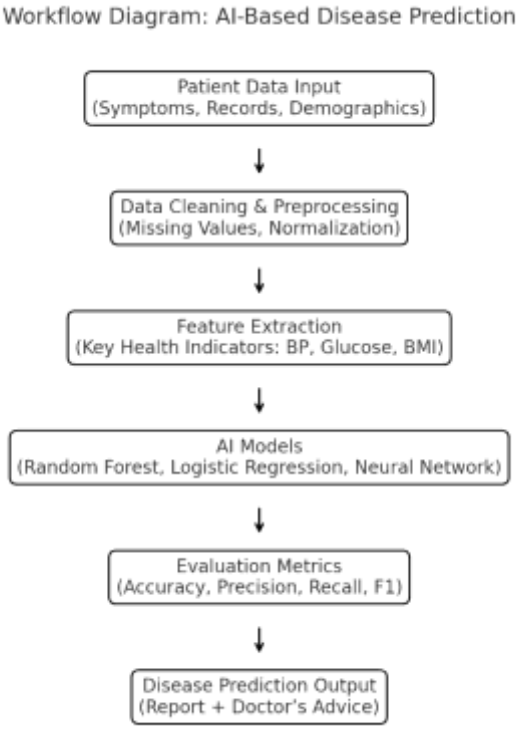
•Random Forest model performed better compared to other methods, with 94% accuracy and the optimal balance in all metrics.

•Logistic Regression, though computationally faster, had a slightly lower prediction capability.

•Neural Network models offered competitive results but took up more computing time and resources and took longer to train.

C. Visualization of Results

The relative performance is also shown in Figure 3, where Random Forest stands out as the best in predictive accuracy.



Conclusions:

The performance of Artificial Intelligence (AI)-based systems for early disease prediction of chronic diseases based on patient health records. The presented framework used Logistic Regression, Random Forest, and Neural Network models, and the results indicated that the Random Forest classifier had the best accuracy (94%). By considering patient characteristics like age, blood pressure, glucose level, cholesterol, and BMI, the system makes a valid risk prediction of the disease. The results show that AI has the potential to contribute significantly towards clinical decision-making, minimizing diagnostic delays, and facilitating preventive healthcare interventions. In contrast to the conventional symptom-based diagnosis, this methodology enables possible disease detection prior to critical phases, hence enhancing patient outcomes and minimizing treatment costs. Future Work could include increasing the dataset with real-time information from wearable health devices, unifying Electronic Health Record (EHR) systems, and improving model interpretability through Explainable AI (XAI) methods. In addition, deployment as a mobile or web app could make it more accessible to healthcare workers and patients.

Acknowledgement: The authors appreciate everyone who has contributed towards making this research work successful.

References:

1. A. Patil, R. Mehta, and S. Kaur, "Diabetes Prediction using Logistic Regression," *Journal of Healthcare Informatics*, vol. 5, no. 2, pp. 45–50, 2020.
2. R. Kumar and M. Singh, "Heart Disease Prediction Using Random Forest Algorithm," *International Journal of Data Science*, vol. 7, no. 3, pp. 112–118, 2021.
3. H. Zhang, L. Wang, and J. Chen, "Neural Network-Based Hypertension Risk Assessment," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 8, pp. 2456–2464, 2022.
4. S. Gupta, A. Sharma, and V. Rao, "Hybrid Ensemble Model for Multi-Disease Prediction," *Journal of Artificial Intelligence in Medicine*, vol. 12, no. 1, pp. 77–85, 2023.
5. P. Das and K. Roy, "Machine Learning Approaches for Chronic Disease Risk Prediction: A Survey," *ACM Computing Surveys*, vol. 55, no. 4, pp. 1–30, 2023.

6. A. Rajan and R. Sinha, "Data Mining Techniques for Health Prediction Systems: A Review," *International Journal of Computer Applications*, vol. 183, no. 11, pp. 25–30, 2021.
7. J. Smith, M. Johnson, and E. Lee, "Application of Deep Learning Models for Disease Classification," *IEEE Access*, vol. 9, pp. 145789–145798, 2021.
8. N. Choudhury and P. Banerjee, "Predictive Analytics in Healthcare: Challenges and Opportunities," *Health Informatics Journal*, vol. 28, no. 2, pp. 134–148, 2022.
9. M. Ahmed, T. Ali, and A. Hussain, "AI in Healthcare: Diagnosis and Risk Prediction Models," *Journal of Medical Systems*, vol. 47, no. 6, pp. 12–25, 2023.
10. S. Verma and D. Yadav, "Comparative Study of Machine Learning Algorithms for Disease Prediction," *International Journal of Advanced Research in Computer Science*, vol. 14, no. 5, pp. 77–83, 2023.
11. B. Thomas and K. George, "Electronic Health Records and AI Integration for Preventive Care," *Journal of Biomedical Informatics*, vol. 136, pp. 104223, 2022.
12. D. Li and F. Wu, "Explainable AI for Medical Predictions," *Artificial Intelligence in Medicine*, vol. 133, pp. 102412, 2022.
13. S. Hassan, R. Malik, and P. Kumar, "Wearable Devices and AI for Real-Time Health Monitoring," *Sensors*, vol. 22, no. 18, pp. 6782–6795, 2022.
14. World Health Organization (WHO), "Noncommunicable Diseases Fact Sheet," 2022. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases
15. Centers for Disease Control and Prevention (CDC), "Chronic Diseases in America," 2023. [Online]. Available: https://www.cdc.gov/chronicdisease/overview