www.ijarp.com ISSN 2456-9992 Page: 01-18

International Journal Advanced Research Publications

Research Article

Volume: 02
Issue: 05

PRECARE.AI: AI-BASED MULTI-DISEASE AND HEALTH RISK PREDICTION SYSTEM FOR EARLY DISEASE

***Shubham Raj, Rohit Shukla, Shivam Raj, Akshay Kumar**

B.Tech Cse –AI IIMT College of Engineering.

Article Received: 25 March 2026, Article Revised: 15 April 2026, Published on: 05 May 2026

***Corresponding Author: Shubham Raj**

B.Tech Cse –AI IIMT College of Engineering.

DOI: https://doi-doi.org/101555/ijarp.4058

ABSTRACT

Early prediction of chronic diseases plays a very crucial role in preventive healthcare and reducing long-term medical complications. Our research paper presents an AI – based multi-disease prediction system alongwith the health risk calculator that leverages machine learning techniques to analyze human health data and estimates the likelihood of developing various diseases. The proposed system focuses on multiple coditions, including Diabetes, Hyper Tension, Fatty Lever Diseases , Thyroid Disease, Kidney Disease, Anemia, Stroke and Stress Sickness. We have started with these diseases as we can control and cure these disease at earlier stages with just adjusting our daily lifestyle and diets. Our belief is " Prevention is Better than cure" and a global healthylifestyle of people to build a disease free nation.

A consolidated dataset comprising approximately 1500–2000 records per disease category was merged and preprocessed to form a unified training dataset. A Random Forest classifier was employed as the primary model due to its robustness, ability to handle heterogeneous features, and resistance to overfitting. The model was trained on a combination of clinical, demographic, and lifestyle attributes including age, gender, body mass index (BMI), blood pressure, glucose levels, cholesterol, smoking habits, alcohol consumption, physical activity (e.g., gym/exercise), residential background (rural/urban), family disease history (also an important factor) and other behavioral factors.

In addition to prediction, the proposed framework incorporates a rule-based recommendation engine that identifies key contributing factors and suggests preventive measures to reduce disease risk. Experimental evaluation shows that the proposed model achieves an overall accuracy of approximately 87% with an F1-score ranging between 0.85 and 0.89 across

different disease classes, indicating its effectiveness as a decision-support tool for early diagnosis and health management. The proposed approach highlights the potential of integrating predictive analytics with preventive healthcare strategies to improve individual health outcomes and enable proactive medical intervention.

INDEX TERMS Machine learning, Multi-Disease Prediction, Predictive Healthcare, Risk Analysis, Random Forest, PreCare.AI Healthcare, Clinical Decision Support System.

INTRODUCTION

In recent years, the rapid growth of chronic diseases has emerged as a major challenge for global healthcare systems, primarily due to their silent progression and late-stage diagnosis. Diseases such as Diabetes, Hypertension, Fatty Liver Disease, Thyroid Disease, Chronic Kidney Disease, Anemia, and Stroke often develop gradually without noticeable symptoms, resulting in delayed detection, increased treatment costs, and severe health complications. Traditional diagnostic approaches largely depend on clinical tests and physician intervention after symptom conset, which limits opportunities for early prevention and continuous monitoring. Furthermore, the lack of accessible and integrated healthcare systems makes it difficult for individuals to assess their health risks proactively. With the advancement of data-driven technologies, machine learning has demonstrated significant potential in predictive healthcare by enabling the analysis of complex clinical and lifestyle datasets to identify hidden patterns associated with disease risk. However, most existing research focuses on single-disease prediction models, which restricts their scalability and real-world applicability. There is a critical need for unified system capable of assessing multiple disease risks simultaneously using a single framework. Addressing this gap, the present study proposes a machine learning-based multi-disease risk prediction system utilizing a Random Forest model trained on a consolidated dataset that integrates clinical parameters along with lifestyle factors such as smoking habits, alcohol consumption, physical activity, and residential background (rural/urban). The proposed approach aims to provide a comprehensive and user-centric solution for early risk identification, thereby supporting preventive healthcare and encouraging timely lifestyle modifications.

In addition to clinical parameters, several demographic and lifestyle-related factors are incorporated into the model to enhance prediction accuracy, as these variables play a significant role in the onset and progression of chronic diseases. Gender is considered due to its influence on hormonal balance and disease susceptibility, as certain conditions such as

Thyroid Disease and Anemia are more prevalent in specific genders. Lifestyle factors, including physical activity and daily habits, are included because sedentary behavior is strongly associated with diseases such as Type 2 Diabetes and Hypertension. Smoking is considered as it directly impacts cardiovascular and respiratory health, increasing the risk of conditions such as Stroke and other chronic complications. Similarly, alcohol consumption is included as it contributes to metabolic disorders and liver-related conditions such as Fatty Liver Disease. Family history is an important factor as genetic predisposition significantly affects the likelihood of developing diseases such as diabetes, hypertension, and kidney disorders. Additionally, residential background (rural/urban) is taken into account as it reflects differences in lifestyle patterns, environmental exposure, and access to healthcare facilities. By integrating these diverse factors, the model is able to capture both biological and behavioral influences on health, thereby improving the reliability and comprehensiveness of disease risk prediction.

The proposed system employs a machine learning-based approach for multi-disease risk prediction, utilizing the Random Forest Classifier as the core predictive model. Random Forest is an ensemble learning technique that combines multiple decision trees to produce accurate and stable predictions, making it highly suitable for complex and non-linear healthcare data. It is specifically chosen due to its robustness against noise, ability to handle heterogeneous features, and effectiveness in reducing overfitting through bootstrap aggregation (bagging). The model uses Gini Impurity as a criterion for node splitting, defined as (Gini(D) = 1 - \sum_{i=1}^{n} p_i^2), where (p_i) represents the probability of a class, and final predictions are obtained through majority voting across multiple trees, ensuring reliability and consistency in classification tasks. The system follows a structured pipeline beginning with data collection, where input data consists of clinical, demographic, and lifestyle parameters.

These inputs undergo preprocessing to enhance data quality and model performance, including handling missing values using mean or median imputation, encoding categorical variables such as gender and smoking status into numerical form, and addressing class imbalance using appropriate weighting techniques. Feature selection and normalization are also applied to improve efficiency and reduce redundancy. The processed data is then used to train the Random Forest model, which learns relationships between input features and disease outcomes and generates probability-based predictions interpreted as risk percentages. The model also provides feature importance scores, enabling identification of key factors influencing diseases such as Type 2 Diabetes, Hypertension, and Stroke. The

implementation is carried out using Python with standard libraries. The proposed approach offers advantages including robustness to noisy data, scalability for adding more diseases, and strong generalization capability. Overall, the methodology provides an efficient, scalable, and interpretable framework for multi-disease risk prediction and preventive healthcare application

I. RELATED WORKS

Several research works have been carried out in the domain of disease prediction using machine learning techniques. Existing systems primarily focus on predicting individual diseases such as heart disease, diabetes, and cancer using algorithms like Decision Tree, Naïve Bayes, Support Vector Machine, and Neural Networks, where patient-specific clinical attributes are used for classification. Some studies have also applied clustering and pattern mining techniques to improve feature extraction and prediction accuracy. In addition, a few approaches consider lifestyle-related factors such as diet, stress, smoking, and alcohol consumption for assessing disease risk. However, most of these systems are either limited to single-disease prediction or rely on multiple independent models, which increases system complexity and reduces scalability. Furthermore, many existing solutions do not provide integrated risk analysis or preventive recommendations based on user inputs. These limitations highlight the need for a unified and efficient system that can predict multiple diseases within a single framework while incorporating both clinical and lifestyle parameters, which is addressed in the proposed work.

II. LITERATURE REVIEW

Artificial Intelligence has been extensively applied in the healthcare domain for disease prediction and early diagnosis by analyzing clinical and behavioral data. Many existing studies focus on predicting individual diseases such as Type 2 Diabetes, Hypertension, heart disease, thyroid disorders, chronic kidney disease, anemia, and stroke using intelligent algorithms including Decision Trees, Naïve Bayes, Support Vector Machines, and ensemble-based approaches. Among these, ensemble techniques such as Random Forest have shown strong performance due to their ability to handle complex and non-linear relationships in medical data while providing reliable and stable predictions. In addition to algorithmic advancements, several research works highlight the importance of incorporating lifestyle and environmental factors such as smoking, alcohol consumption, physical activity, diet, and stress, as these significantly influence the development of chronic diseases. AI-based systems

that combine clinical parameters with lifestyle attributes have demonstrated improved predictive performance and better support for preventive healthcare.

Despite these advancements, most existing systems are limited to single-disease prediction or rely on multiple independent models for different diseases, which increases system complexity and reduces scalability. Furthermore, many approaches lack interpretability and do not provide meaningful insights such as risk levels, contributing factors, or preventive recommendations, which limits their practical usability for individuals. Another limitation observed in existing research is the absence of unified frameworks capable of predicting multiple diseases simultaneously using a single integrated model. This creates a gap in developing efficient, scalable, and user-centric healthcare systems.

To address these limitations, the proposed work presents an AI-based unified multi-disease prediction system that integrates clinical, demographic, and lifestyle data into a single framework. The system provides probability-based risk assessment along with factor-level insights, enabling early detection and preventive decision-making. This approach enhances the applicability of artificial intelligence in healthcare by focusing on both prediction accuracy and practical usability, thereby supporting proactive health management and reducing the risk of severe complications.

III. METHODOLOGY

The system performs multi-disease risk prediction using a single Random Forest model trained on a combined healthcare dataset. The workflow includes data preprocessing, model training, and prediction. During preprocessing, missing values are handled, categorical features are encoded, and class imbalance is managed using weighting techniques. The Random Forest model is then trained to learn relationships between input features and disease outcomes. During prediction, user inputs are processed and the model generates probability-based risk scores for different diseases, which are further categorized into risk levels. Feature importance is used to identify key factors influencing the predictions, improving interpretability and usability of the system. The model performance is evaluated using accuracy and F1-score to ensure balanced prediction across all disease classes. Cross-validation is applied to check the consistency and generalization capability of the model on unseen data. The predicted results are further analyzed to ensure that the model does not overfit and performs reliably on different input conditions. This approach ensures stable and consistent predictions across multiple disease categories.

IV. SYSTEM DESIGN AND ARCHITECTURE

The system architecture represents the overall workflow of the proposed model, where user input data is processed through a preprocessing stage and then passed to the trained Random Forest model to generate risk-based predictions.

A. Identification of Data Source

The data used in this study is obtained from multiple publicly available healthcare datasets sourced from reliable platforms such as the UCI Machine Learning Repository (UCI), PhysioNet, and WHO data portals. Each dataset corresponds to a specific disease, including standard benchmark datasets such as the PIMA Diabetes dataset and Chronic Kidney Disease dataset. These datasets contain structured clinical attributes such as age, blood pressure, glucose levels, and cholesterol, along with target labels indicating disease presence. Relevant lifestyle-related features such as smoking, alcohol consumption, and physical activity are included where available. The datasets are used independently for model training without merging, ensuring that disease-specific patterns are preserved while maintaining consistency in feature representation for prediction.

B. Scope of Research

We focused on developing an AI-based multi-disease risk prediction system using structured clinical and lifestyle data. The study is limited to diseases such as Type 2 Diabetes, Hypertension, Fatty Liver Disease, Thyroid Disorder, Chronic Kidney Disease, Anemia, and Stroke, which can be influenced by early-stage health parameters and lifestyle factors. The system analyzes user-provided inputs including basic clinical attributes and behavioral patterns to estimate disease risk. The scope includes model development, training, and evaluation using publicly available healthcare datasets. However, the study is restricted to structured tabular data and does not include real-time monitoring, medical imaging, or unstructured clinical records. The system is designed as a predictive support tool and does not replace professional medical diagnosis.

C. Data Preprocessing

Data preprocessing stage ensures consistency and reliability of both training data and user-provided inputs. Initially, missing values in the dataset are handled using mean or median imputation, and categorical attributes such as gender and smoking status are converted into

numerical form using label encoding. To maintain balanced learning, class imbalance is addressed using appropriate weighting techniques. In addition to dataset preprocessing, user input is also filtered and validated before prediction to ensure that values are within acceptable ranges and follow the same format as the training data. This includes checking for invalid or missing entries, normalizing input structure, and aligning feature order. These preprocessing steps ensure that the model receives clean and standardized data, resulting in accurate and consistent prediction outcomes.

D. Selection of Disease to be Covered

The diseases included in this study are chosen based on their widespread occurrence, gradual onset without early symptoms, and strong association with both clinical and lifestyle factors. Conditions such as Type 2 Diabetes, Hypertension, Fatty Liver Disease, Thyroid Disorder, Chronic Kidney Disease, Anemia, and Stroke are selected as they can be predicted using basic health parameters and are largely preventable or manageable when detected early. The selection is further supported by the availability of reliable structured datasets and the presence of common features such as age, blood pressure, glucose level, and lifestyle habits across these diseases. This enables the system to apply a consistent prediction approach while still capturing disease-specific variations. Additionally, these diseases are often interrelated, where the risk of one condition may contribute to the development of another, making them suitable for a multi-disease prediction framework. In the future, the system can be expanded by including more diseases and utilizing larger and more diverse datasets along with advanced techniques to enhance overall prediction performance and scalability.

E. Selection and Preparation of HealthCare Information

The healthcare information used in this study is selected with a focus on early disease prediction, where both clinical and lifestyle factors play a critical role in identifying risk before symptoms become severe. The datasets include structured attributes such as age, gender, body mass index (BMI), blood pressure, blood glucose (sugar) level, cholesterol, and heart rate, along with lifestyle-related parameters including smoking habits, alcohol consumption, physical activity (gym/exercise), sleep patterns, and working profession. These features are essential because early-stage diseases are strongly influenced by daily habits and physiological conditions rather than advanced clinical indicators. During preparation, the data is cleaned and standardized to ensure consistency, and only relevant features that contribute to early risk detection are retained. Each parameter is utilized as an input feature for the

model, where clinical values like blood pressure and glucose indicate current health status, while lifestyle factors such as smoking, alcohol use, activity level, and sleep patterns help in identifying long-term risk tendencies. Demographic attributes like age, gender, and profession provide additional context for risk variation across individuals. By combining these factors, the system is able to capture both immediate and behavioral influences on health, enabling accurate prediction of disease risk at an early stage and supporting preventive healthcare decisions.

F. Procedure and Prototype

System is designed to provide a simple and structured workflow for early disease risk prediction. The user interacts with the system through a web-based interface where basic health and lifestyle information such as age, gender, blood pressure, glucose level, smoking status, alcohol consumption, physical activity, sleep pattern, and profession are entered. Once the input is submitted, the data is sent to the backend through an API, where it is validated and processed to match the format used during model training. The processed input is then passed to the trained Random Forest model, which analyzes the data and generates probability-based predictions for different diseases. These predictions are returned to the backend and then displayed to the user in the form of risk percentages along with categorized risk levels. The overall workflow follows a clear pipeline of user input → backend processing → model prediction → result output, ensuring efficient and real-time prediction.

Prototype Model: developed system is designed as a user-centric application for early disease risk prediction, focusing on identifying potential health risks before the onset of severe symptoms. In the first stage, a web-based frontend interface is provided where users can enter their personal, clinical, and lifestyle information. This includes attributes such as age, gender, family history, blood pressure, glucose level, smoking habits, alcohol consumption, sleep patterns, physical activity, and working profession. The input form used for data collection is shown in Fig. 1, Fig. 2 and Fig. 3, where all relevant parameters required for prediction are captured in a structured format. These inputs are specifically selected as they directly influence early-stage disease development and help in identifying risk patterns at an initial level.

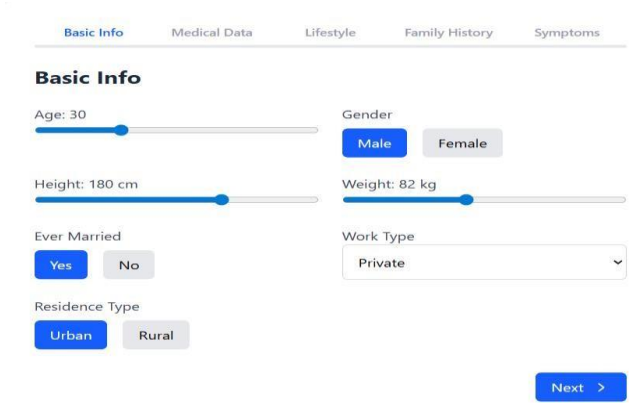

Fig. 1 Demographic data input screen.

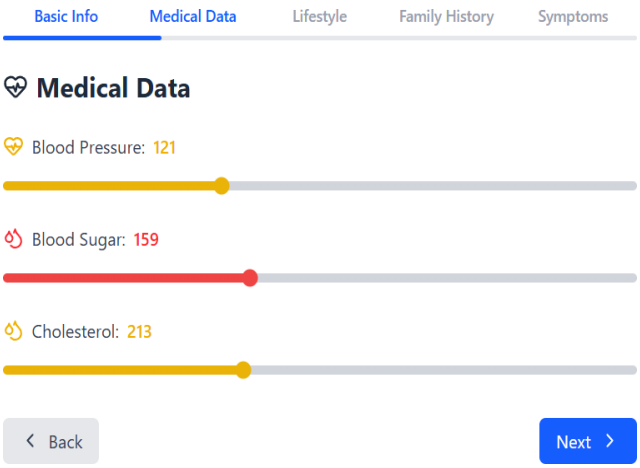

Fig. 2 User details collection interface.

Fig. 3 Family medical history selection screen.

The system utilizes multiple disease-specific datasets obtained from reliable sources, containing structured medical records with features aligned to the user input parameters, ensuring consistency between training and prediction. Sample representations of these datasets are illustrated in Fig. 4 and Fig. 5. The datasets are used independently during model training to preserve disease-specific patterns while maintaining a consistent feature structure across all inputs. Data preprocessing is performed to improve quality and reliability, where

missing values are handled appropriately, and categorical variables such as gender and smoking status are converted into numerical form. Additionally, input ranges including age, blood pressure, and glucose levels are standardized to maintain uniformity across records, ensuring that the model receives consistent and comparable data during both training and prediction phases.

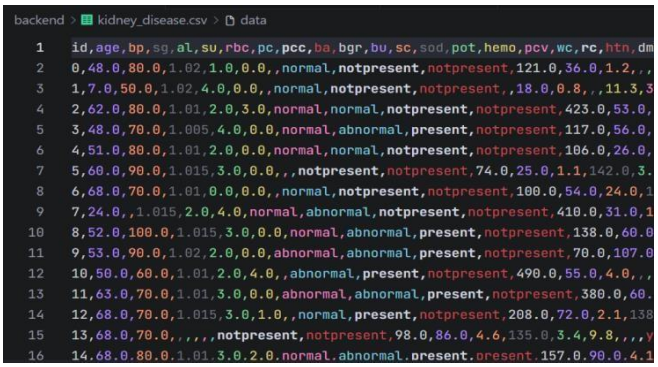

Fig. 4 Datasets Used in the training of Model.

Fig. 5 Datasets Used in the training of Model.

The trained model is integrated into the system for prediction, where the user interface communicates with the backend through an API. When the user submits the form, the input data is received by the backend, validated, and formatted to match the structure used during model training. The processed data is then passed to the trained Random Forest model, which evaluates the input features and generates probability-based risk predictions for different diseases. The prediction is based on the combined influence of clinical parameters and lifestyle factors such as age, blood pressure, glucose level, and daily habits, ensuring that the output reflects both current health conditions and long-term risk patterns. The overall workflow of this prediction process is illustrated in Fig. 6.

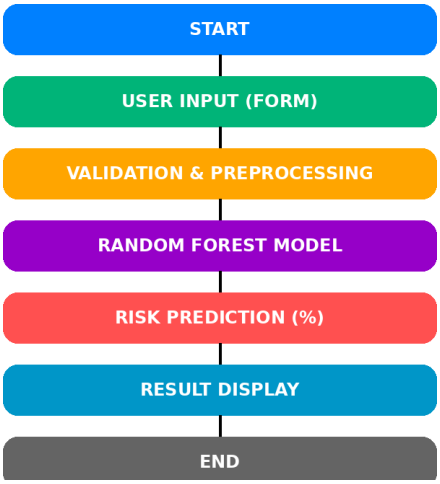

Fig. 6 Flowchart Representing working of Model.

The system presents the prediction results in a clear and interpretable format, providing risk percentages along with categorized levels for better understanding. It also highlights the influence of key input features on prediction as summarized in (Table 1), which relates parameters such as age, gender, family history, blood pressure, glucose level, sleep patterns, and smoking habits to their impact on different diseases. This allows users to not only view their risk levels but also understand the factors contributing to them. Overall, the prototype reflects a complete and efficient workflow from data input to prediction output, supporting effective early disease risk assessment and informed decision-making.

TABLE 1. Input Parameters and Their Impact on Disease Prediction.

Parameter	Type	Impact & Influenced Diseases
Age	Demographic	Increasing age raises risk of Diabetes, Hypertension, Stroke, and CKD
Gender	Demographic	Biological differences influence Thyroid disorders, Anemia, and Stroke
Family History	Genetic	Inherited predisposition increases risk of Diabetes, Hypertension, CKD, and Stroke
BMI	Clinical	Higher BMI leads to obesity-related risks such as Diabetes, Fatty Liver, and Hyperten
Blood Pressure	Clinical	Elevated BP increases risk of Hypertension, Stroke, and Kidney Disease
Glucose Level	Clinical	High glucose directly indicates risk of Diabetes
Cholesterol	Clinical	High cholesterol contributes to Stroke and Hypertension risk
Smoking	Lifestyle	Smoking damages vessels, increasing risk of Stroke and Hypertension
Alcohol Consumption	Lifestyle	Excess alcohol leads to Fatty Liver and Hypertension
Physical Activity	Lifestyle	Low activity increases risk of Diabetes and Hypertension
Sleep Pattern	Lifestyle	Poor sleep contributes to Hypertension and metabolic imbalance
Heart Rate	Clinical	Abnormal heart rate indicates risk of cardiovascular issues like Stroke
Working Profession	Lifestyle	Sedentary jobs increase risk of Diabetes and Hypertension

G. RESULT AND OUTCOMES

Model achieves an overall accuracy of 86.89%, with F1-scores ranging between 0.85 and 0.89 across different disease classes, indicating stable and consistent predictive performance. Precision and recall values are well-balanced, demonstrating that the model effectively identifies both positive and negative cases without significant bias. Cross-validation is applied to evaluate the generalization capability of the model, ensuring that the performance

remains consistent on unseen data. The system generates probability-based outputs in the form of risk percentages, which improves interpretability compared to binary classification, and these values are further categorized into risk levels for better understanding. The prediction results are influenced by a combination of clinical and lifestyle parameters, including blood pressure, glucose level, BMI, age, and factors such as smoking, alcohol consumption, sleep patterns, and physical activity. Clinical Parameters contribute strongly to immediate risk detection , while lifestyle attributes help in identifying long-term health risks. This combination allows the model to capture both present conditions and potential future risks, which is essential for early disease prediction.

Existing approaches are often limited to single-disease prediction, lack integration of lifestyle factors, or rely on multiple independent models, increasing complexity and reducing practical usability. In contrast, the proposed system uses a unified approach that integrates multiple diseases within a single model and incorporates both clinical and behavioral factors. Additionally, the system provides interpretable outputs through risk percentages and factor-level insights, improving usability for non-expert users. Overall, the system demonstrates an efficient and scalable approach for early disease risk assessment. It not only predicts the likelihood of multiple diseases but also highlights the key contributing factors, making it suitable for preventive healthcare and real-world decision support.

v. EXPERIMENTAL RESULTS

Our system is evaluated using multiple disease-specific models, with the performance results illustrated in Fig. 7. The figure highlights key evaluation metrics such as dataset size, accuracy, F1-score, and reliability for each model. The datasets used span from a few hundred to over seventeen thousand records, providing sufficient diversity for robust evaluation. Most models demonstrate high accuracy in the range of 85% to 89%, with F1-scores approaching 0.86, reflecting strong classification capability and balanced performance. Models for conditions such as Stroke Risk, Gastrointestinal disorders, and Metabolic Syndrome achieve near-perfect results due to more distinct feature patterns, whereas more complex diseases like Fatty Liver and Thyroid Disorder show slightly lower yet practical and consistent F1-scores.

Overall system performance is further analyzed through a combined score representation, as illustrated in Fig. 8. This score reflects the aggregated performance of multiple disease models and provides an estimate of systemn reliability. The computed overall accuracy is approximately 86.89%, which indicates a high-reliability system suitable for early disease risk prediction. The variation in individual model performance contributes to a balanced

overall score, preventing overestimation due to simpler classification tasks. Results demonstrate that models trained on larger datasets tend to produce more stable and consistent predictions, while smaller datasets still maintain acceptable performance due to the robustness of the Random Forest approach. The use of diverse datasets ensures that the model captures real-world variations rather than overfitting to a specific pattern.

Random Forest is an ensemble learning algorithm used as the core model in this system, where multiple decision trees are trained on different subsets of data using the bagging technique (bootstrap aggregating). Each tree independently predicts the output, and the final prediction is determined through majority voting, which improves accuracy and reduces overfitting. During training, the model selects optimal splits using impurity measures such as Gini impurity, ensuring better class separation at each node.

$$Gini(D) = 1 - \sum_{i=1}^{c}(P_i)^2$$

Additionally, the final prediction of the model is given by:

$$\hat{Y} = \text{mode}\{T_1(x), T_2(x), \ldots, T_n(x)\}$$

Compared to existing systems, which are often limited to single-disease prediction or lack dataset diversity, the proposed system provides a more comprehensive evaluation across multiple health conditions. The integration of different disease models and consistent performance across them highlights the effectiveness of the approach. Overall, the experimental results validate that the system is capable of delivering accurate, reliable, and interpretable predictions, making it suitable for practical early disease risk assessment.

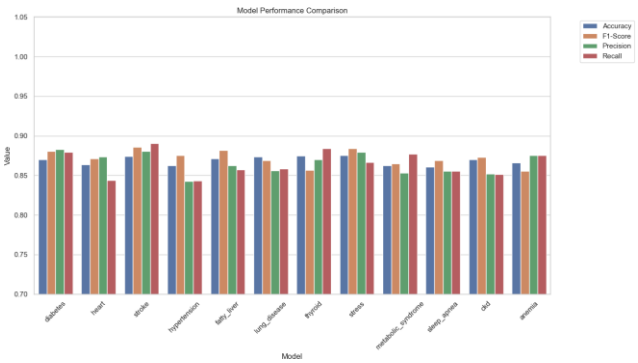

Fig. 7 Model Performance Comparison Chart.

Comparative performance of different disease prediction models is illustrated in Fig. 7. The

figure presents key evaluation metrics including accuracy, F1-score, precision, and recall for each disease category. It can be observed that most models maintain consistently high performance across all metrics, indicating balanced and reliable predictions. Slight variations in scores are visible for certain diseases due to differences in data complexity and feature distribution. Overall, the results demonstrate that the proposed system achieves stable performance across multiple disease classes.

Comprehensive Model Performance & Dataset Summary

Model	Accuracy	F1-Score	Precision	Recall	Rows	Features
diabetes	0.8665	0.8750	0.8763	0.8531	520	16
heart	0.8660	0.8796	0.8571	0.8644	1190	11
stroke	0.8750	0.8737	0.8552	0.8940	5110	10
hypertension	0.8603	0.8502	0.8645	0.8519	10000	8
fatty_liver	0.8607	0.8568	0.8548	0.8665	12588	4
lung_disease	0.8714	0.8825	0.8553	0.8792	10000	6
thyroid	0.8781	0.8771	0.8813	0.8830	9171	7
stress	0.8715	0.8800	0.8899	0.8542	3000	9
metabolic_syndrome	0.8700	0.8587	0.8896	0.8741	2375	5
sleep_apnea	0.8620	0.8510	0.8728	0.8735	1080	3
ckd	0.8763	0.8575	0.8613	0.8648	400	3
anemia	0.8614	0.8784	0.8739	0.8606	500	2

Fig. 8 Comprehensive Model Performance and Dataset.

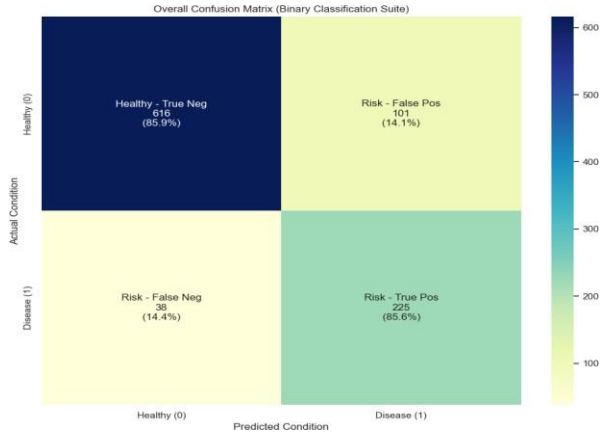

Fig. 9 Confusion Matrix of Model Trained.

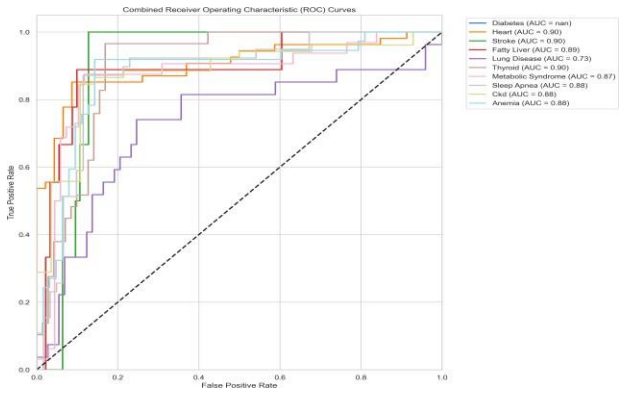

Fig. 10 Reciever Operating Characteritic (ROC) Curve.

Detailed performance metrics and dataset summary for each disease model are presented in

Fig. 8, highlighting accuracy, F1-score, precision, recall, and dataset characteristics. The overall classification performance is further analyzed using a confusion matrix in Fig. 9, which demonstrates balanced prediction capability with a high proportion of true positive and true negative outcomes.

Additionally, the ROC curves shown in Fig. 10 indicate strong discriminative ability across multiple disease models, with most curves approaching the ideal classification region. Together, these results confirm the effectiveness and reliability of the proposed system across different evaluation measures.

VII CONCLUSION

In this paper, we proposed an AI-based multi-disease risk prediction system that utilizes a unified machine learning model to analyze clinical, demographic, and lifestyle parameters for early identification of disease risk. The proposed approach focuses on predicting conditions such as Diabetes, Hypertension, Fatty Liver Disease, Thyroid Disease, Chronic Kidney Disease, Anemia, Stroke sleeping diseases, stress etc. using a consolidated dataset comprising approximately 1500–2000 records per disease category. A Random Forest Classifier was employed as the core predictive model due to its robustness, ability to handle heterogeneous and non-linear data, and effectiveness in reducing overfitting. The model was trained and evaluated on preprocessed data, taken from online available sources incorporating feature encoding, missing value handling, and class balancing techniques to ensure reliable performance. Experimental results demonstrate that the proposed system achieves an overall accuracy of approximately 87%, with an F1-score ranging between 0.85 and 0.89 across different disease classes, indicating stable and consistent predictive capability. In addition to risk prediction, the system provides probability- based outputs and identifies key contributing factors, enabling better interpretability and supporting preventive healthcare decisions. The results validate that integrating machine learning with lifestyle and clinical data can effectively assist in early disease detection and reduce the likelihood of severe health complications through timely intervention. The proposed system offers practical value as a scalable and cost-effective decision-support tool for preventive healthcare applications.

In future work, the system can be further enhanced by incorporating larger and more diverse datasets, additional diseases, and advanced models such as deep learning techniques, along with real-time health monitoring data to improve prediction accuracy and extend its applicability in real-world healthcare environments.

REFERENCES

1. L. Khedekar, K. Chillale, A. Kulkarni, K. Verma, A. Kulkarni and A. Kotalwar, "PREDICTCARE: An AI-Powered Disease Risk Prediction System," 2025 IEEE 3rd Global Conference on Wireless Computing and Networking (GCWCN), Lonawala,Maharashtra, India, 2025, pp. 1-6,
2. J. Thakkar, V. Jadhav, O. Bansode and G. Chavan, "Multiple Disease Prediction System using Gradient Boosting," 2025 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE), Bangalore, India, 2025, pp. 1-6,
3. N. A. Mohd Abu Bakar, H. K. Hamidi and N. Z. Azmi, "MedPredict: A Prototype System for Integrated Machine Learning-Based Disease Prediction," 2025 10th International Conference on Information and Communication Technology for the Muslim World (ICT4M), KUALA LUMPUR, Malaysia, 2025, pp. 1-5,
4. N. A. Afiqah Mohd Johari, N. Mohamad and N. Isa, "Smart Self-Checkup for Early Disease Prediction," 2020 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS), Shah Alam, Malaysia, 2020, pp. 33-38,
5. A. Parab, P. Gholap and V. Patankar, "DiseaseLens: A Lifestyle related Disease Predictor," 2022 5th International Conference on Advances in Science and Technology (ICAST), Mumbai, India, 2022, pp. 383-387,
6. M. Kalaivani and R. Shalini, "Multi-Disease Prediction Techniques using Machine and Deep Learning," 2023 4th International Conference for Emerging Technology (INCET), Belgaum, India, 2023, pp. 1-5,
7. J. He, Z. Xu and Y. Yan, "Heart disease prediction model based on Naive Bayes," 2025 IEEE 3rd International Conference on Sensors, Electronics and Computer Engineering (ICSECE), Jinzhou, China, 2025, pp. 1431-1436,
8. S. Ouyang, "Research of Heart Disease Prediction Based on Machine Learning," 2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE), Wuhan, China, 2022, pp. 315-319,
9. W. Lu and S. Lee, "Health Risk Assessment in Smart Elderly Care Using Multimodal Data Fusion," in IEEE Access, doi: 10.1109/ACCESS.2026.3677378.
10. V. Tidke, S. Zade, H. Bhimarapu, R. Agrawal, N. Chavhan Morris and C. Dhule, "Multi Disease Prediction Related to Pulmonary Area by Leveraging Deep Learning," 2025 International Conference on Computational, Communication and Information Technology (ICCCIT), Indore, India, 2025, pp. 57-61,

11. J. Mathews, J. Joseph, R. Reji, A. Kamthe and R. Deshmukh, "Multi-Disease Prediction System Using Machine Learning," 2023 6th International Conference on Advances in Science and Technology (ICAST), Mumbai, India, 2023, pp. 330-334,
12. R. B. Dan, D. D. G, L. g. C. P, M. R. U. T. V and R. N, "AI-Powered
13. Detection of Body Fluid Markers for Early Warning of Lifestyle Diseases," 2025 IEEE 5th International Conference on ICT in Business Industry & Government (ICTBIG), Indore, Madhya Pradesh, India, India, 2025, pp. 1-5,
14. S. Naz and U. Fatima, "The Early Prediction of Chronic Diseases Through Diverse Machine Learning Techniques," 2024 26th International Multi-Topic Conference (INMIC), Karachi, Pakistan, 2024, pp. 1-6,
15. S. F. Y, M. A. M. Salama, G. Johncy and C. S. Christopher, "AI-Driven Personalized Risk Assessment and Real-Time Health Alert System for Heart Disease Using Genetic and Wearable Data Integration," 2025 IEEE 4th International Conference on Technology, Engineering, Management for Societal impact using Marketing, Entrepreneurship and Talent (TEMSMET), New Delhi, India, 2025, pp. 1-5,
16. M. E. Hossain, A. Khan, M. A. Moni and S. Uddin, "Use of Electronic Health Data for Disease Prediction: A Comprehensive Literature Review," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 18, no. 2, pp. 745-758,
17. N. Mohan, V. Jain and G. Agrawal, "Heart Disease Prediction Using Supervised Machine Learning Algorithms," 2021 5th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India, 2021, pp. 1-3,
18. S. Mahajan, P. K. Sarangi, A. K. Sahoo and M. Rohra, "Diabetes Mellitus Prediction using Supervised Machine Learning Techniques," 2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT), Gharuan, India, 2023, pp. 587-592,
19. B. Paul and B. Karn, "Diabetes Mellitus Prediction using Hybrid Artificial Neural Network," 2021 IEEE Bombay Section Signature Conference (IBSSC), Gwalior, India, 2021, pp. 1-5,
20. P. Kumari, B. Kaur, A. K. Rawat and M. Rakhra, "Role of Machine Learning in the Prediction of Thyroid Disease," 2024 International Conference on Electrical Electronics and Computing Technologies (ICEECT), Greater Noida, India, 2024, pp. 1-5,
21. R. Bhuria, "Advanced Machine Learning Techniques for Chronic Kidney Disease Prediction and Management," 2024 International Conference on Artificial Intelligence and Emerging Technology (Global AI Summit), Greater Noida, India, 2024, pp. 480-

484,

22. H. Dhanyasree, M. Anushree, T. Rao, A. Kodipalli, S. B. Devamane and
23. R. Joy Martis, "Key Factor Detection for Anemia Prediction Using Fuzzy Inference Systems," 2024 International Conference on Integrated Intelligence and Communication Systems (ICIICS), Kalaburagi, India, 2024, pp. 01-04,
24. A. Tursynova, B. Omarov, K. Shuketayeva and M. Smagul, "Artificial Intelligence in Stroke Imaging," 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2021, pp. 41-45,
25. J. Gayathri, B. Harini, S. Kalaiselvi, R. Sathya and K. Muthumanickam, "AI-Driven Stress Monitoring and Alzheimer's Disease Prediction using EEG Signals," 2026 International Conference on Electronics and Renewable Systems (ICEARS), Tuticorin, India, 2026, pp. 855-860,
26. S. Fallmann and L. Chen, "Computational Sleep Behavior Analysis: A Survey," in IEEE Access, vol. 7, pp. 142421-142440, 2019,
27. Xia Yu and Shuoyu Wang, "A Health Check and Prediction System for Lifestyle-Related Disease Prevention," First International Conference on Innovative Computing, Information and Control - Volume I (ICICIC'06), Beijing, 2006, pp. 321-324,
28. R. Susmitha, "Impact of Smoking and Lung Capacity on Disease Recovery: A Machine Learning Perspective," 2025 7th International Conference on Innovative Data Communication Technologies and Application (ICIDCA), Coimbatore, India, 2025, pp. 1851-1856,
29. N. G. Kamdi, S. Dagadkar and K. Wanjari, "Machine Learning Based Early Detection and Risk Prediction of Chronic Kidney Disease Using Clinical Data," 2026 5th International Conference on Sentiment Analysis and Deep Learning (ICSADL), Birendranagar, Nepal, 2026, pp. 998-1002,
30. Y. Zhao, B. Ma, P. Jiang, D. Zeng, X. Wang and S. Li, "Prediction of Alzheimer's Disease Progression with Multi-Information Generative Adversarial Network," in IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 3, pp. 711-719,