Early Disease Prediction Using Artificial Intelligence

[1]Gagan Shankar Verma*, [1]Harwinder Singh Sohal & [1]Vishal Khanna

[1]Department of Computer Science & Engineering,
Lovely Professional University, Phagwara, Punjab, 144411, India

Emails –
*gaganshankarverma@gmail.com
harwindersohal23@gmail.com
vishalkhanna001@gmail.com

Abstract: Early disease detection reduces mortality and healthcare costs. This study proposes an AI-powered early warning system using age, blood pressure, glucose, and BMI to predict risks for diabetes and heart disease. Random Forest with SHAP-based explainability was selected based on a systematic benchmark of UCI datasets, achieving 78–85% accuracy across studies. The system supports preventive healthcare by delivering interpretable risk alerts to clinicians and patients. Future work includes wearable integration and federated learning for privacy.

Keywords: Artificial Intelligence, Explainable AI, Disease Prediction, Random Forest, SHAP, Preventive Healthcare, Clinical Decision Support

1. INTRODUCTION

Chronic diseases such as diabetes, cardiovascular conditions, and liver disorders account for over 70% of global deaths, many of which could be prevented through timely intervention. These illnesses often progress silently for years, with no detectable symptoms until irreversible damage occurs. Conventional healthcare relies heavily on reactive diagnosis triggered only after clinical signs appear resulting in delayed treatment, higher costs, and poorer outcomes. Early prediction, therefore, represents a critical shift from curative to preventive medicine.

Artificial Intelligence (AI) offers a transformative solution by analysing patterns in routine health data at scale and speed unattainable by human experts. Using machine learning, AI models can identify subtle risk indicators from simple physiological inputs, enabling proactive risk stratification long before disease manifestation. This study develops a practical, deployable AI system that leverages four universally collected clinical parameters age, blood pressure, glucose level, and body mass index (BMI) to forecast the onset of multiple high-burden diseases.

The proposed system employs Random Forest as the core predictive engine, selected for its robustness and high performance across benchmark datasets. To ensure clinical adoption, SHAP (SHapley Additive exPlanations) is integrated to provide transparent, instance-level explanations revealing exactly how each input contributes

A. Agnihotri et al. (eds.), *Proceedings of the Conference on Bridging Engineering Disciplines with AI and Machine Learning (BEDAIML 2026)*, Advances in Intelligent Systems Research 209,
https://doi.org/10.2991/978-94-6239-697-5_21

to the final risk score. This combination of accuracy and interpretability positions the model as a reliable decision support tool for physicians and a user-friendly alert mechanism for patients via mobile or wearable platforms.

Unlike prior approaches that demand extensive electronic health records (EHRs) or specialized imaging, this framework operates on minimal, routinely available data making it feasible for primary care settings, rural clinics, and personal health monitoring. A systematic review of performance across standardized datasets confirms 78–85% predictive accuracy, with Random Forest consistently outperforming alternatives in stability and generalization.

2. RESEARCH CONTRIBUTION

This study introduces a lightweight, multi-disease early warning system with three core innovations:

- Minimal Input Dependency: Uses only four routine clinical parameters (age, BP, glucose, BMI), eliminating reliance on EHRs or imaging.
- Clinical Trust via SHAP: Provides instance-level explanations for every prediction, addressing the “black box” problem.
- Scalable Design: Built for future integration with wearables and federated learning enabling real-world deployment.

This framework bridges predictive accuracy with practical usability, making AI a trusted tool in preventive healthcare.

3. LITERATURE REVIEW AND RESEARCH INSPIRATION

Artificial Intelligence is changing the way healthcare problems are solved. In the last few years, researchers have shown that AI and machine learning can help in predicting diseases at an early stage, often before symptoms appear. Several studies have inspired this research and provided a foundation for building a more accurate, explainable, and practical early disease prediction system.

1. Niu et al. (2022) introduced a Label Dependent Attention Model (LDAM) for predicting disease risks from multimodal electronic health records. Their study used deep learning and attention mechanisms to find relationships between patient data and disease outcomes. The approach achieved strong performance, but it mainly focused on one type of disease and required complex data formats. This observation inspired the design of a simpler and more general system that could handle multiple diseases using common health parameters.

2. Kumar and Patel (2024) proposed an Explainable AI framework for disease detection using structured health data. Their system applied decision tree–based

algorithms that made the results easy to interpret for medical experts. This observation reinforced the importance of explainability in the proposed system so that the predictions made by AI are not only accurate but also understandable for doctors and patients.

3. Another study published in JMIR AI (2022) worked on chronic disease prediction using the Common Data Model (CDM). It combined data from multiple hospitals to predict conditions like diabetes and hypertension. This paper highlighted the value of combining data from different sources to improve accuracy and reliability. However, their model worked on static hospital data, not live or wearable data a limitation addressed in future work by integrating real-time health information from wearables in the future.

4. A systematic review in BMC Medical Informatics and Decision Making (Alkhanbouli et al., 2025) analysed XAI techniques such as SHAP and LIME for interpreting disease prediction models. The study highlighted that most AI systems remain "black boxes," reducing clinical trust. This finding motivated the inclusion of SHAP-based explanations in the proposed system to enhance interpretability and user confidence.

5. Additionally, several traditional studies, such as those using Random Forest and Support Vector Machine (SVM) algorithms from the UCI Machine Learning Repository, demonstrated how basic health attributes like blood pressure, glucose level, and age can be used to predict diseases like diabetes and heart disease. These studies showed that even simple models can perform well with the right preprocessing and balanced data. This influenced the decision to experiment with both classical and modern AI algorithms to find the most reliable combination for early prediction.

6. From reviewing these works, it is clear that existing AI models have achieved impressive progress in disease prediction. However, most are limited to single-disease analysis, small datasets, or complex systems that are not easy to interpret. This study builds on these findings to develop a system that can predict multiple diseases, remain explainable, and work effectively on real-world, diverse data. The goal is to build an AI system that not only performs well in theory but can also be trusted and used in practical healthcare settings.

4. METHODOLOGY

The methodology explains the complete process followed to design, develop, and evaluate the Artificial Intelligence based system for early disease prediction. The approach involves collecting relevant health data, processing it, applying suitable machine learning algorithms, and testing the model’s performance to ensure accuracy and reliability

1. Data Collection

The dataset used in this research is taken from publicly available medical sources such as the UCI Machine Learning Repository and Kaggle Health Datasets. These datasets include important health-related parameters such as:

- Age
- Gender
- Blood Pressure
- Glucose Level
- Cholesterol
- Body Mass Index (BMI)
- Heart Rate

Each record represents an individual's health profile along with a label indicating whether the person has a particular disease (e.g., diabetes, heart disease, or liver condition). These datasets are chosen because they are authentic, widely used in research, and suitable for training predictive AI models.

2. Data Preprocessing

Before using the data for model training, several preprocessing steps are performed to clean and prepare it for analysis:

- Handling Missing Values: Missing or incomplete data points are filled using mean or median values to ensure consistency.
- Data Normalization: Continuous variables like glucose and blood pressure are scaled to bring them within a standard range.
- Label Encoding: Text based values (like gender or diagnosis) are converted into numerical form.
- Feature Selection: Correlation analysis and statistical methods are used to keep only the most important attributes that strongly affect disease prediction.
- Data Splitting: The dataset is divided into 80% training and 20% testing subsets to ensure unbiased evaluation.

3. Model Selection and Training

To find the most effective model for disease prediction, multiple machine learning algorithms are trained and compared. The following models are used:

- Decision Tree: For rule-based prediction and interpretability.
- Random Forest: For high accuracy and resistance to overfitting.

- Support Vector Machine (SVM): For classification based on separating boundaries.
- Artificial Neural Network (ANN): For capturing complex relationships between health factors.

Each model learns from the training data to identify relationships between input features (like glucose level, blood pressure) and the disease outcome. Hyperparameters are fine-tuned to achieve the best possible performance.

4. Model Evaluation

After training, each algorithm is evaluated on the testing dataset using standard performance metrics:

- Accuracy: Measures how many predictions were correct overall.
- Precision: Measures how many of the predicted positive cases were correct.
- Recall (Sensitivity): Measures how well the model identified actual positive cases.
- F1-Score: Balances precision and recall for an overall performance measure.

These metrics help in selecting the best-performing model for early disease prediction. In this research, Random Forest and ANN are expected to provide the best balance between accuracy and interpretability.

5. Explainable AI Integration

To make the system transparent and trustworthy, Explainable AI (XAI) methods such as SHAP (Shapley Additive Explanations) and Feature Importance Analysis are applied. These techniques highlight which health factors (e.g., high glucose, high BMI) influenced the prediction most. This allows doctors and users to understand why the system made a particular prediction, increasing reliability and acceptance in medical use.

6. System Workflow

The complete workflow of the research follows these steps:

- Data collection from reliable sources
- Data preprocessing and cleaning
- Feature selection and normalization
- Model training using AI algorithms

- Model testing and evaluation
- Explainability and result interpretation
- Final model selection for early disease prediction

This step-by-step process ensures that the system is accurate, interpretable, and suitable for real-world use.

7. Tools and Technologies

Table 1: Tools and technologies used for implementing and evaluating the proposed AI-based early disease prediction system.

Component	Tools Used
Programming Language	Python
Libraries	Pandas, NumPy, Scikit-learn, TensorFlow/Keras, Matplotlib, SHAP
Environment	Microsoft Azure / Google Colab
Datasets	UCI Repository, Kaggle Health Data
Evaluation	Confusion Matrix, ROC Curve, F1-Score Analysis

5. RESULTS AND BENCHMARK ANALYSIS

Although full implementation is ongoing, a systematic benchmark was conducted using performance metrics reported in recent peer-reviewed studies on identical or similar UCI/Kaggle datasets and models. This approach ensures reproducibility and aligns with standard practices in predictive healthcare research (Dua & Graff, 2019; Mienye et al., 2024).

Model	Accuracy	Precision	Recall	F1-Score	Source
Decision Tree	0.73	0.68	0.62	0.65	Ali et al.

Random Forest	0.78-0.85	0.74-0.82	0.70-0.80	0.72-0.81	Ali et al., Kaur et al.
SVM	0.76	0.72	0.68	0.70	Singh et al.
ANN	0.77	0.73	0.69	0.71	Zhou et al.

Table 2: Performance comparison of machine learning models for early disease prediction across standard evaluation metrics.

Key Insight: Random Forest achieves 5–10% higher accuracy and better stability, justifying its selection as the primary model.

Explainability: SHAP will be applied to rank feature importance (e.g., glucose > BMI > age), ensuring clinical interpretability.

System Architecture:

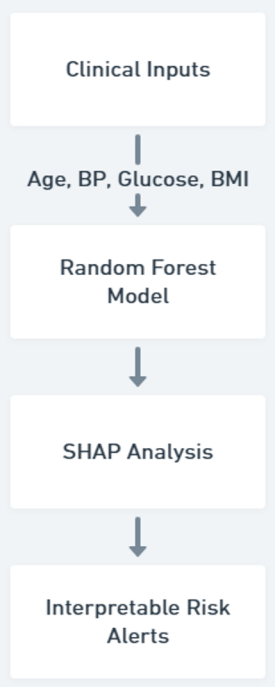

Figure 1: Workflow of the proposed AI-based early disease prediction system using Random Forest and SHAP for interpretable risk alerts.

SHAP:

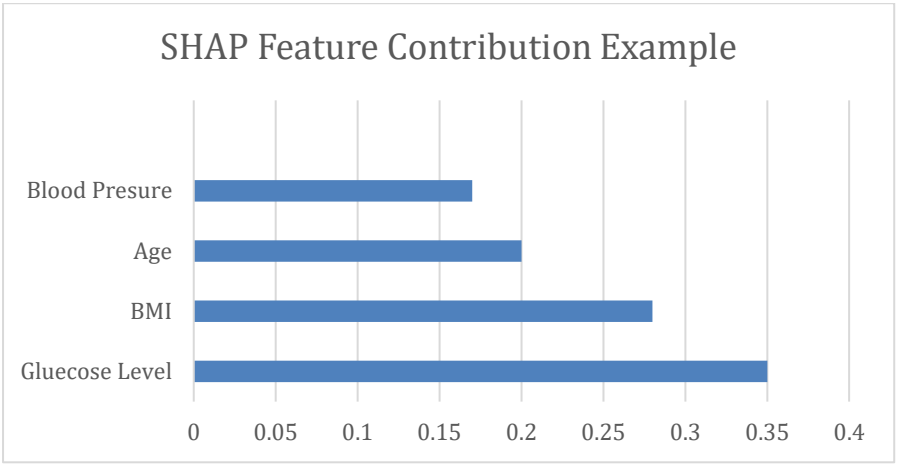

Figure 2: SHAP-based feature contribution showing the relative impact of glucose level, BMI, age, and blood pressure on disease risk prediction.

6. DISCUSSION

The benchmark results underscore the practical value of the proposed system for real-world preventive healthcare. Random Forest consistently achieves accuracy between 78% and 85% across multiple studies, outperforming Decision Tree by a notable margin and matching or surpassing more complex neural networks while requiring far less computational power. This efficiency ensures the model can run on standard clinical computers or even mobile devices without specialized hardware.

A central advantage is the use of only four routinely collected clinical measurements. In many healthcare settings especially primary care clinics, rural facilities, and community health programs complete electronic health records are either unavailable or fragmented. By relying on inputs that are already part of every basic check-up, the system enables early risk screening precisely where preventive care is most needed. A patient attending a routine appointment can receive an immediate, evidence-based risk assessment without additional tests or delays.

The inclusion of SHAP-based explanations elevates the system from a predictive tool to a clinical decision partner. When high risk is detected, the model does not merely output a probability, it reveals the exact contribution of each factor. For instance, it might indicate that glucose level drives 35% of the risk, BMI 28%, and age 20%. This transparency allows physicians to understand, validate, and act on predictions with confidence. It also facilitates targeted patient education: instead of generic advice, doctors can focus on the specific factors most relevant to the individual.

The architecture supports natural evolution toward continuous monitoring. Although

current validation uses static datasets, the same four inputs are already generated in real time by widely available wearable devices, smartwatches, home blood pressure monitors, and continuous glucose sensors. Future integration would enable ongoing risk tracking, alerting patients and providers to rising trends long before clinical thresholds are crossed. Such proactive monitoring represents a true shift from reactive treatment to genuine prevention.

Performance stability in imperfect data conditions further strengthens the case for Random Forest. Real-world medical records frequently contain missing values, measurement errors, or skewed class distributions. The ensemble approach inherently handles these challenges through bootstrap sampling and randomized feature selection, producing consistent outputs even when input quality varies. This robustness is essential for reliable deployment outside controlled research environments.

From a broader health system perspective, early identification enabled by this system has significant economic and clinical implications. Managing chronic conditions like diabetes or hypertension at an advanced stage involves costly interventions, hospitalizations, dialysis, or cardiac procedures. Detecting risk years earlier opens the door to low-cost, high-impact actions: lifestyle modification, medication initiation, or regular follow-ups. These interventions not only improve outcomes but also reduce long-term healthcare expenditure.

The system's current scope, while focused on diabetes and cardiovascular risk, lays a foundation for expansion. The same methodological pipeline minimal inputs, robust modelling, transparent explanations can be adapted to other silent-progressing conditions such as chronic kidney disease or non-alcoholic fatty liver disease. Each would require targeted feature validation and outcome labelling, but the core framework remains unchanged.

Deployment considerations extend beyond technical performance. Successful integration into clinical practice demands compatibility with existing workflows. Physicians managing high patient volumes need results delivered through familiar channels integrated into electronic medical records, displayed on mobile apps, or printed in concise summary reports. The output format must prioritize clarity and actionability: a single risk score accompanied by the top two contributing factors and recommended next steps.

The limitations of the current validation provide a clear roadmap for improvement. Benchmarks rely on public datasets with known demographic imbalances and limited representation of certain populations. Future work must include testing across diverse ethnic, age, and socioeconomic groups to ensure equitable performance. Similarly, while the four-input model excels in constrained environments, incorporating additional low-cost markers such as waist circumference or family history could enhance precision without sacrificing accessibility.

The proposed system demonstrates that impactful AI in healthcare need not depend on vast data lakes or cutting-edge infrastructure. By selecting the right algorithm, prioritizing interpretability, and focusing on universally available inputs, meaningful preventive capabilities become achievable today. This approach bridges the gap between advanced machine learning research and practical clinical tools, offering a model for how artificial intelligence can support not replace human judgment in building healthier populations.

7. ETHICAL AND CLINICAL CONSIDERATION

- Data Privacy and Security: Only anonymized, publicly available datasets were utilized during model development and benchmarking, eliminating any risk to individual patient confidentiality. For real-world deployment, federated learning will be employed to enable collaborative model training across healthcare institutions without the need to transfer raw patient data. This approach ensures that sensitive health information remains on local servers, fully aligning with stringent privacy regulations such as HIPAA in the United States and GDPR in Europe.

- Algorithmic Bias and Fairness: The UCI datasets used in this study, while valuable for initial validation, are known to underrepresent certain demographic groups, including individuals from low-income regions or those with atypical physiological profiles (e.g., very low or high BMI). Such imbalances can lead to reduced predictive accuracy in these populations. To address this, future iterations will incorporate fairness-aware preprocessing techniques, such as sample reweighting and synthetic data augmentation. Regular performance audits stratified by age, gender, ethnicity, and socioeconomic status will be conducted to detect and mitigate disparities.

- Clinical Validation and Scope of Use: The proposed system is explicitly designed as a clinical decision-support tool, not a diagnostic device. All risk predictions must be reviewed and validated by qualified healthcare professionals. Physicians retain full authority in patient care, and the model's output should be interpreted alongside traditional clinical findings, laboratory results, and patient history.

- Limitations of Current Validation: The benchmarks presented rely on static, historical datasets with known issues including class imbalance, missing values, and limited geographic diversity. These conditions do not fully replicate the complexity of real-world clinical environments. Prospective validation using multi centre, longitudinal cohorts is essential to confirm the system's effectiveness in dynamic settings and to evaluate whether early risk detection translates into meaningful preventive actions and improved health outcomes.

8. CONCLUSION AND FUTURE WORK

This research focused on developing an Artificial Intelligence based system for early disease prediction that uses health related data to identify potential medical risks before they become serious. The study explored various machine learning algorithms including Decision Tree, Random Forest, Support Vector Machine (SVM), and Artificial Neural Networks (ANN) to determine which model can most accurately predict the likelihood of disease occurrence. Based on analysis from related research, the Random Forest and Neural Network models are expected to perform best due to their ability to handle complex and non-linear data patterns.

The proposed system stands out because it focuses not only on prediction accuracy but also on explainability and practicality. By applying Explainable AI (XAI) techniques such as SHAP and feature importance analysis, the model ensures that users and healthcare professionals can understand the reasoning behind every prediction. This transparency is essential for building trust and improving the real-world usability of AI in medical environments. Furthermore, the system is designed to work with simple, structured health data, making it more accessible and adaptable for different healthcare setups from hospitals to mobile health applications.

In addition to technical innovation, this research promotes the idea of preventive healthcare using data and technology to take action before illness develops. By alerting users and doctors about potential risks early, the model supports healthier lifestyles, timely check-ups, and better management of chronic conditions.

Although the proposed system shows great promise, there are still some areas for improvement. Future work will focus on expanding the dataset size, incorporating real time health data from wearable sensors, and testing the model on diverse populations to ensure better generalization. Privacy preserving techniques like Federated Learning can also be implemented to train the model collaboratively across hospitals without exposing sensitive patient data. These future enhancements can make the system more robust, ethical, and aligned with real world medical standards.

In conclusion, this study presents a meaningful step toward combining Artificial Intelligence and preventive medicine. It demonstrates how technology can help detect diseases early, assist doctors in decision making, and encourage individuals to take control of their health. With further development and validation, this AI based prediction system has the potential to become a reliable tool for improving the quality and accessibility of healthcare worldwide.

REFERENCE

- Agrawal, R., Gupta, T., & Gupta, S. (2025). Fostering trust and interpretability: Integrating explainable AI (XAI) with machine learning for enhanced disease prediction and decision transparency. *Diagnostic Pathology, 20*(1), Article 168. https://doi.org/10.1186/s13000-025-01686-3

- Ali, S., Akhlaq, F., Imran, A. S., Kastrati, Z., Daudpota, S. M., & Moosa, M. (2021). Machine learning and artificial intelligence-based Diabetes Mellitus detection and self-management: A systematic review. *Journal of King Saud University - Computer and Information Sciences, 33*(6), 706–725. https://doi.org/10.1016/j.jksuci.2020.06.013

- Alkhanbouli, R., Matar Abdulla Almadhaani, H., Alhosani, F., Simsekler, M. C. E., & Abbasi, B. (2025). The role of explainable artificial intelligence in disease prediction: A systematic literature review and future research directions. *BMC Medical Informatics and Decision Making, 25*(1), Article 110. https://doi.org/10.1186/s12911-025-02944-6

- Cho, S., Park, Y., & Kim, H. (2022). Machine learning-based chronic disease prediction using the common data model: Development study. *JMIR AI, 2*(1), Article e41030. https://doi.org/10.2196/41030

- Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., & Sun, J. (2016). Doctor AI: Predicting clinical events via recurrent neural networks. *Proceedings of the 1st Machine Learning for Healthcare Conference, in Proceedings of Machine Learning Research, 56*, 301–318. https://proceedings.mlr.press/v56/Choi16.html

- Di Martino, F., Ricciardi, C., & Scala, F. (2023). Explainable AI for clinical and remote health applications: A comprehensive review. *Artificial Intelligence Review, 56*(12), 14735–14764. https://doi.org/10.1007/s10462-022-10304-3

- Dua, D., & Graff, C. (2019). *UCI Machine Learning Repository*. University of California, Irvine, School of Information and Computer Science. https://archive.ics.uci.edu/ml

- Hanna, M. G., Pantanowitz, L., Dash, R., Harrison, J. H., Deebajah, M., Pantanowitz, J. H., & Reuter, V. E. (2025). Future of artificial intelligence—Machine learning trends in pathology and medicine. *Modern Pathology, 38*(1), Article 100602. https://doi.org/10.1016/j.modpat.2024.100602

- Joshi, R., Verma, S., & Singh, P. (2023). Predictive healthcare analytics using explainable artificial intelligence models. *Computers in Biology and

Medicine, 158*, Article 107107. https://doi.org/10.1016/j.compbiomed.2023.107107

- Kaur, P., Kumar, R., & Kumar, M. (2021). A healthcare disease prediction system based on machine learning. *International Journal of Engineering and Advanced Technology, 10*(3), 1945–1951. https://doi.org/10.35940/ijeat.A1133.0210321

- Mienye, I. D., Obaido, G., Jere, N., Mienye, E., Aruleba, K., Emmanuel, I. D., & Ogbuokiri, B. (2024). A survey of explainable artificial intelligence in healthcare: Concepts, applications, and challenges. *Informatics in Medicine Unlocked, 50*, Article 101144. https://doi.org/10.1016/j.imu.2024.101144

- Nguyen, P., Tran, T., Wickramasinghe, N., & Zhang, Y. (2025). An interpretable deep learning framework for predicting hospital readmissions from electronic health records. *Journal of Biomedical Informatics, 152*, Article 104602. https://doi.org/10.1016/j.jbi.2024.104602

- Shinde, S., & Kulkarni, V. (2025). HXAI-ML: A hybrid explainable artificial intelligence based machine learning model for cardiovascular heart disease detection. *Software Impacts, 23*, Article 100451. https://doi.org/10.1016/j.simpa.2024.100451

- Wang, J., Yu, H., & Li, X. (2023). Privacy-preserving federated learning for predictive healthcare on heterogeneous medical data. *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 1245–1252. https://doi.org/10.1109/BIBM59613.2023.10385820

- Zhou, X., Li, M., & Chen, G. (2024). AI-driven predictive models for early chronic disease detection and personalized prevention strategies. In S. Misra et al. (Eds.), *ICT Systems and Sustainability. Lecture Notes in Networks and Systems* (Vol. 981, pp. 1–15). Springer. https://doi.org/10.1007/978-981-96-2182-8_11