An Intelligent Machine Learning-Based Multi-Disease Prediction System with Chatbot

Dr. A.V.S Siva Rama Rao
Department of CSE – AIML
Sasi Institute of technology and Engineering

[1] **Ananthapalli Uday Sai**
Department of CSE – AIML
Sasi Institute of technology and Engineering
udaysri.ananthapalli@sasi.ac.in

[2] **Kadiyam Amaravani**
Department of CSE – AIML
Sasi Institute of technology and Engineering
amaravani.kadiyam @sasi.ac.in

[3] **Chinta Ravi Kumar**
Department of CSE – AIML
Sasi Institute of technology and Engineering
ravikumar.chinta@sasi.ac.in

[4] **Katta Naga Sai**
Department of CSE – AIML
Sasi Institute of technology and Engineering
nagasai.katta@sasi.ac.in

Abstract

Early diagnosis of chronic diseases is crucial to improve healthcare results and mitigate medical risk. In the present paper, an intelligent system of machine learning-based multi-disease prediction with the integration of chatbot technology for healthcare support is proposed. The system predicts the risk of **Diabetes**, **Parkinson's** disease, and **Lung Cancer** using machine learning and deep learning models. The system uses the **XGBoost** algorithm to predict Diabetes and Lung Cancer by analyzing structured medical data, while the **EfficientNet** CNN is used to predict Parkinson's by analyzing handwriting images.

The system is implemented using **Python** with libraries such as **TensorFlow**, **Keras**, and **Scikit-learn**. A chatbot module is integrated to provide basic health guidance and explain prediction results. Experimental results show that the proposed system achieves reliable prediction accuracy and supports early disease screening. The proposed approach demonstrates the potential of artificial intelligence to assist in accessible and intelligent healthcare decision support systems.

KEYWORDS: MACHINE LEARNING, MULTI-DISEASE PREDICTION, XGBOOST, EFFICIENTNET, HEALTHCARE AI, MEDICAL DATA ANALYSIS, DISEASE RISK PREDICTION, INTELLIGENT CHATBOT.

1. Introduction

Artificial intelligence and machine learning are widely used in healthcare for early disease detection and clinical decision support, helping to reduce health risks and improve treatment outcomes [19], [21]. Early diagnosis is especially important for diseases such as Diabetes, Parkinson's disease, and Lung cancer, where timely intervention can significantly improve patient outcomes.

Machine learning techniques, particularly XGBoost, have demonstrated high accuracy in predicting diabetes by analyzing clinical and lifestyle data [1], [3], [9]. Similarly, deep learning approaches have been applied for early detection of Parkinson's disease through handwriting pattern analysis, providing a non-invasive diagnostic method [4], [15]. In the case of lung cancer, both XGBoost-based models and convolutional neural networks have shown strong performance in predicting disease risk and analyzing medical imaging data [6], [7], [26].

In this paper, an intelligent multi-disease prediction system is proposed using machine learning and deep learning techniques. The system predicts Diabetes and Lung cancer using the XGBoost algorithm and detects Parkinson's disease using EfficientNet based on handwriting images. It is implemented using Python and integrates a chatbot to provide basic health information and explain prediction results, improving user interaction and accessibility. This integrated approach aligns with recent advancements in deep learning and healthcare systems, making disease prediction more efficient and accessible [22], [24], [25].

2. Literature Survey

Recent studies have explored machine learning and deep learning techniques for early disease prediction using medical data. These models can analyze healthcare datasets and identify patterns that assist in detecting diseases at early stages. Various algorithms have been applied to predict diseases such as Diabetes, Parkinson's disease, and Lung cancer. The following studies highlight key contributions in this area.

Wang et al. (2020) proposed a diabetes prediction model using the XGBoost algorithm to analyze clinical and lifestyle attributes of patients. The study compared XGBoost with traditional machine learning algorithms such as Support Vector Machine and Random Forest. Experimental results showed that the XGBoost model achieved higher prediction accuracy and improved classification performance for diabetes risk prediction. However, the study relied on a relatively small dataset, which may affect the generalization capability of the model [1].

Kavakiotis et al. (2017) conducted a comprehensive review of machine learning and data mining methods used in diabetes research. Their study analyzed various supervised learning techniques applied to healthcare datasets and reported that models such as decision trees, neural networks, and support vector machines are widely used for predicting diabetes risk. However, the authors noted that many studies are limited by dataset imbalance and lack of diverse clinical features [2].

Gundogdu (2023) developed a diabetes prediction system using the XGBoost classifier with Random Forest–based feature selection techniques. The proposed approach improved prediction efficiency by selecting the most relevant medical attributes from the dataset. Experimental results demonstrated strong classification performance for diabetes detection. However, the system requires proper feature engineering to achieve optimal results [3].

Razaq et al. (2025) proposed a deep learning framework for detecting Parkinson's disease using handwriting patterns such as spiral and wave drawings. The authors used the EfficientNet convolutional neural network to extract meaningful features from the handwriting images. The model achieved high prediction accuracy for Parkinson's disease detection. However, deep learning models require large datasets and higher computational resources for effective training [4].

Benredjem et al. (2024) introduced a multimodal prediction framework that combines handwriting data and clinical information to improve Parkinson's disease detection. The study applied deep learning techniques to capture complex neurological patterns from multiple data sources. The results showed improved prediction performance compared with single-data approaches. However, integrating multimodal datasets increases system complexity and training time [5].

Lin et al. (2023) developed a machine learning model for predicting Lung cancer risk using metabolic biomarkers and clinical attributes. The authors implemented the XGBoost algorithm to analyze medical datasets and identify significant risk factors associated with lung cancer. Experimental results indicated that the model can support early screening and diagnosis of lung cancer. However, the study emphasized the need for larger and more diverse clinical datasets for improving prediction reliability [6].

Das et al. (2025) proposed a deep learning–based system for detecting Lung cancer using convolutional neural networks and medical imaging data. Their approach focused on extracting significant features from CT scan images to identify cancerous regions. The experimental results showed improved classification accuracy compared to traditional machine learning models. However, the model requires large annotated medical image datasets and significant computational resources for effective training [7].

Various algorithms have been applied to predict diseases such as diabetes, Parkinson's disease, and lung cancer. Although previous studies show that machine learning and deep learning techniques can predict diseases most systems focus on a single disease and use limited datasets. Therefore, a unified system capable of predicting multiple diseases is needed. The proposed system addresses this gap by integrating multiple prediction models with a chatbot for early disease screening.

Table-1: Comparison of Existing Disease Prediction Methods

Authors & Year	Model Architecture	Dataset Used	Performance	Result	Limitations
Wang et al. 2020 [1]	XGBoost	Clinical Diabetes Dataset	Accuracy: 89%	Effective diabetes risk prediction	Small dataset size
Kavakiotis et al. 2017 [2]	Multiple ML Models	Diabetes Healthcare Data	Comparative evaluation	Identified effective ML methods	Dataset imbalance
Gundogdu 2023 [3]	XGBoost + Random Forest Feature Selection	Sylhet Diabetes Dataset	Accuracy: 95.6%	Improved prediction efficiency	Requires feature engineering
Razaq et al. 2025 [4]	EfficientNet CNN	Spiral & Wave Handwriting Dataset	Accuracy: 95.3%	Effective Parkinson detection	High computational cost
Benredjem et al. 2024 [5]	Multimodal Deep Learning	Handwriting + Clinical Data	Accuracy: 96%	Improved neurological disorder prediction	Model complexity
Lin et al. 2023 [6]	XGBoost	Lung Cancer Biomarker Dataset	Accuracy: 75%	Supports early lung cancer screening	Requires larger datasets
Das et al. 2025 [7]	CNN	Lung CT Image Dataset	High classification accuracy	Effective cancer detection	Requires large annotated images

3. Analysis of Datasets

The performance of the proposed multi-disease prediction system was evaluated by utilizing various healthcare datasets. The data was collected from various publicly available sources, including Kaggle. The datasets used were medical data regarding various diseases, including Diabetes, Parkinson's disease, and Lung cancer. The data includes clinical attributes, patient symptoms, medical images, etc. This helps machine learning models learn patterns regarding various medical conditions.

Unlike many previous research works, which used a single dataset for model training, the proposed prediction system utilizes data regarding various medical conditions. This increases data diversity. Therefore, the proposed system is able to provide accurate predictions regarding various medical conditions. The proposed prediction system is more efficient compared to previous research works.

The proposed prediction system utilizes a diabetes dataset, which includes various patient health attributes, including age, BMI, blood glucose level, hypertension, smoking habits, etc. For Parkinson's disease prediction, a handwriting dataset was used. The dataset includes spiral and wave drawings. The proposed prediction system also utilizes a lung cancer dataset, which includes medical symptoms and lifestyle habits.

After preprocessing the data, the datasets were split into training sets and test sets. Approximately 80% of the data was used to train the models, while the remaining 20% was used to test the models.

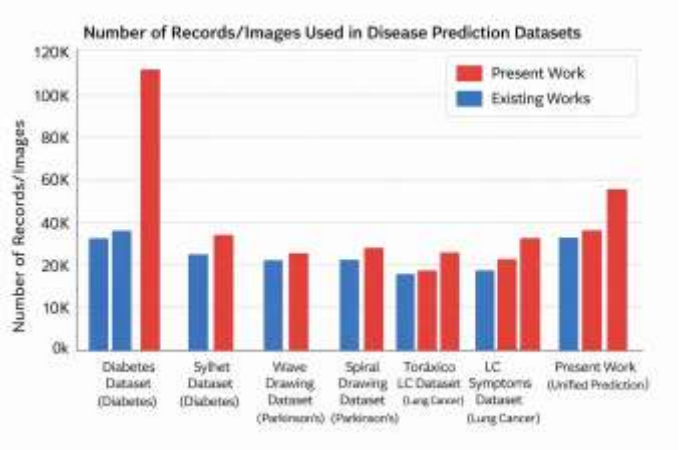

Fig. 1: Dataset distribution used for the proposed multi-disease prediction system.

4. Methodology of Proposed System

The proposed system is designed to predict multiple diseases using machine learning and deep learning techniques. The system analyzes medical data and images to detect diseases such as Diabetes, Parkinson's disease, and Lung cancer. The main objective of the system is to provide an intelligent healthcare platform that assists users in early disease prediction and improves awareness about potential health risks.

Initially, the user provides input data through a web-based interface. For diabetes and lung cancer prediction, the system accepts structured medical attributes such as age, BMI, smoking history, blood glucose level, hypertension status, and other health-related indicators. For Parkinson's disease detection, the system accepts handwriting images such as spiral and wave drawings, which are commonly used in neurological examinations to identify tremor patterns associated with the disease.

Before training the models, the input data undergoes several preprocessing steps. In the case of structured medical data, preprocessing includes handling missing values, encoding categorical variables into numerical form, and applying feature scaling techniques to normalize the dataset. These preprocessing steps help improve the accuracy and stability of the machine learning models. For image data, preprocessing includes resizing the images to a fixed dimension, converting them into numerical arrays, and normalizing pixel values to prepare them for deep learning analysis.

After preprocessing, the datasets are used to train machine learning and deep learning models. The diabetes and lung cancer prediction modules use the XGBoost classification algorithm, which is an ensemble learning technique based on gradient boosting decision trees. This algorithm is effective for structured healthcare datasets and helps identify complex relationships between medical attributes. For Parkinson's disease detection, the system uses the EfficientNet convolutional neural network architecture to extract deep visual features from handwriting images and classify them into healthy or diseased categories.

Once the models are trained, they are integrated into the system for real-time prediction. When the user submits input data, the system processes the data through the trained models and generates prediction results indicating the likelihood of disease. The results are displayed to the user through the interface. Additionally, an intelligent chatbot module is incorporated to provide explanations of prediction results and offer basic health guidance, making the system more interactive and user-friendly.

4.1 Gradient Boosting Principle

Gradient boosting had been an ensemble learning technique where weak prediction models had been combined to produce a stronger model. Each new model had attempted to correct the errors made by the previous models.
The predicted output of the model had been defined as:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)$$

Formula: 1

Where:

- $\hat{y}_i$ = predicted output
- f_k = individual decision tree
- K = total number of trees
- x_i = input feature vector

Each decision tree had contributed to the final prediction.

Objective Function

The objective function of XGBoost had consisted of two components:

$$L(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

Formula: 2

Where:

- $l(y_i, \hat{y}_i)$= training loss function
- $\Omega(f_k)$= regularization term
- y_i= actual output
- $\hat{y}_i$ = predicted output

The regularization term had controlled model complexity and prevented overfitting.

Regularization Term

The regularization term had been defined as:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \| w \|^2$$

Formula: 3

Where:

- T= number of leaves
- w= leaf weights
- γ= complexity parameter
- λ= regularization parameter

Regularization had ensured that the model remained generalized.

4.1 System Architecture

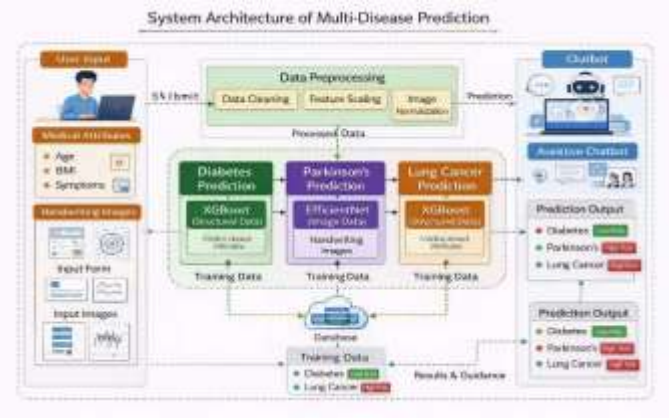

Fig. 2: Architecture of the Intelligent Multi-Disease Prediction System

The figure illustrates the architecture of the proposed intelligent healthcare system designed for predicting multiple diseases using machine learning and deep learning techniques. The architecture consists of several interconnected modules that process user inputs, perform data preprocessing, apply prediction models, and generate the final output results.

Initially, the system receives input data from the user through a web interface. The input may include medical attributes such as age, body mass index (BMI), blood glucose level, smoking history, and other health indicators for predicting diseases such as Diabetes and Lung cancer. For detecting Parkinson's disease, the system accepts handwriting images such as spiral and wave drawings which are commonly used in neurological tests.

After receiving the input, the data is passed to the preprocessing module. In this stage, the system performs operations such as data cleaning, normalization, feature encoding, and image resizing. These preprocessing steps ensure that the data is transformed into a suitable format for machine learning and deep learning models.

Once preprocessing is completed, the processed data is forwarded to the disease prediction module. In this module, the diabetes and lung cancer prediction tasks are performed using the XGBoost machine learning algorithm, which is effective for analyzing structured healthcare datasets and identifying complex relationships between medical attributes.

For Parkinson's disease detection, the system uses the EfficientNet deep learning architecture. This convolutional neural network analyzes handwriting images and extracts important visual features such as tremor patterns and irregular drawing movements that may indicate the presence of Parkinson's disease.
Each prediction module generates a probability score indicating whether the patient is likely to have the disease. These outputs are then processed by the result generation module, which produces the final prediction result.

Finally, the prediction results are displayed to the user through the web interface. The system also integrates an intelligent chatbot module that provides

explanations and basic healthcare guidance based on the prediction results. This architecture enables the system to efficiently analyze multiple types of medical data and provide reliable predictions within a unified healthcare platform.

4.2 Block Diagram

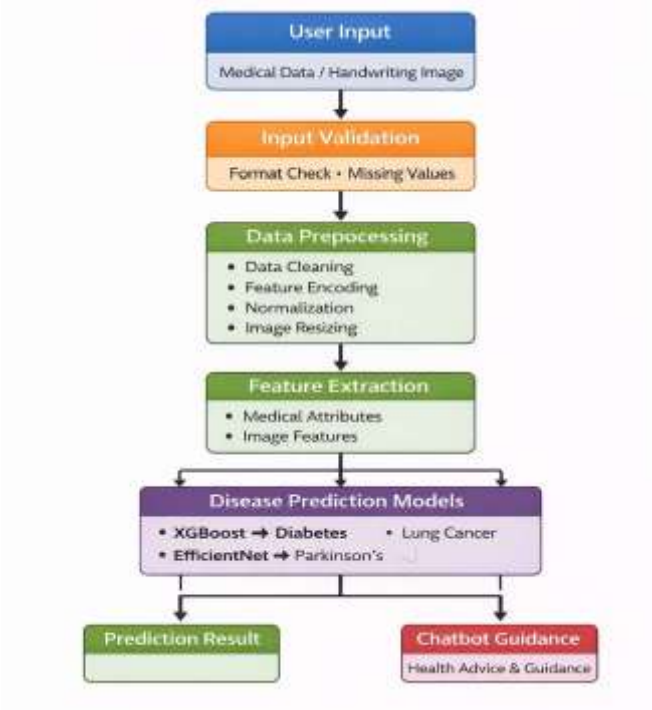

Fig. 3: Block Diagram of the Proposed Multi-Disease Prediction System

The block diagram illustrates the overall workflow of the proposed intelligent healthcare prediction system. The process begins with the **user input module**, where the user provides medical information or uploads the required data through the system interface. The system accepts structured medical attributes for predicting diseases such as Diabetes and Lung Cancer. In addition, handwriting images such as spiral or wave drawings are used for detecting Parkinson's disease.

After receiving the input, the data is passed to the **input validation module**, where the system verifies the format, completeness, and correctness of the entered data. This stage ensures that all input values are valid and suitable for further processing.

The validated data is then forwarded to the **data preprocessing module**. In this stage, several preprocessing operations are performed, including data cleaning, feature encoding, normalization, and image resizing. These preprocessing steps help convert the raw input data into a structured format suitable for machine learning and deep learning models.

After preprocessing, the processed data is passed to the **feature extraction and prediction module**.

In this module, structured medical data is analyzed using the **XGBoost machine learning algorithm**, which is effective for handling structured healthcare datasets and identifying important predictive patterns. For Parkinson's disease detection, the system uses the **EfficientNet deep learning model** to extract meaningful visual features from handwriting images.

Each prediction model generates probability scores indicating the likelihood of disease occurrence. These outputs are then processed by the **result aggregation module**, which combines the prediction results and determines the final outcome.

Finally, the system displays the prediction results through the **output module**, where users can view the predicted disease risk. Additionally, an integrated chatbot module provides basic healthcare guidance and explanations based on the prediction results, making the system more interactive and user-friendly.

5. Implementation

The proposed multi-disease prediction system was implemented using machine learning and deep learning frameworks combined with a web-based interface to enable user interaction. The implementation was carried out using Python programming language along with various libraries for data processing and model development. The system integrates algorithms for predicting diseases such as Diabetes, Parkinson's disease, and Lung cancer.

Initially, the datasets required for training the models were collected from publicly available sources. These datasets contain structured healthcare attributes and medical information used for disease prediction. Before training the models, data preprocessing steps such as handling missing values, feature encoding, normalization, and image resizing were performed to ensure consistent data representation.

The machine learning model for diabetes and lung cancer prediction was implemented using the XGBoost algorithm, which is effective for analyzing structured medical datasets.

For Parkinson's disease detection, the system uses the EfficientNet deep learning architecture to analyze handwriting images and extract relevant visual features.

The system was developed using several supporting technologies including Python libraries such as NumPy, Pandas, Scikit-learn, TensorFlow, and Keras for data processing and model training. A web-based interface was developed using the Flask framework to allow users to provide input data and receive prediction results in real time.

During the training phase, the dataset was divided into training and testing sets to evaluate model performance. The trained models were then integrated into the web application so that users can enter medical information and obtain prediction results through the interface. The final system provides disease prediction results along with basic health guidance through an integrated chatbot module.

6. Experimental Results

The performance of the proposed multi-disease prediction system was evaluated using machine learning and deep learning models for predicting diseases such as Diabetes, Parkinson's disease, and Lung Cancer. The models were trained using the prepared datasets and evaluated using standard performance metrics such as accuracy, precision, recall, and F1-score.

The diabetes and lung cancer prediction modules were implemented using the **XGBoost algorithm**, which is effective for analyzing structured medical datasets. The Parkinson's disease detection module was implemented using the **EfficientNet deep learning architecture**, which extracts visual features from handwriting images such as spiral and wave drawings.

To evaluate the classification performance of the models, confusion matrices and standard evaluation

metrics were used. These metrics help analyze how accurately the system identifies disease cases and normal cases.

Accuracy Calculation

Accuracy measures the proportion of correctly classified samples among all samples.

The **formula** for accuracy is:

Formula 4:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

Where:

- **TP (True Positive)** – correctly predicted disease cases
- **TN (True Negative)** – correctly predicted healthy cases
- **FP (False Positive)** – healthy cases incorrectly predicted as disease
- **FN (False Negative)** – disease cases incorrectly predicted as healthy

Higher accuracy indicates better performance of the prediction model.

6.1 Precision, Recall, and F1-Score

Precision, recall, and F1-score are used to evaluate the classification quality of the model.

Precision

Formula 5:

$$\text{Precision} = TP / (TP + FP)$$

Precision measures how many predicted positive cases are actually correct.

Recall

Formula 6:

$$\text{Recall} = TP / (TP + FN)$$

Recall measures how many actual disease cases are correctly identified by the model.

F1-Score

Formula 7:

$$\text{F1 Score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$$

The F1-score balances both precision and recall and provides a better measure when dealing with imbalanced datasets.

6.3 Model Performance Evaluation

The experimental results show that the proposed system achieves reliable performance in predicting multiple diseases. The XGBoost model demonstrates strong performance in analyzing structured medical datasets for diabetes and lung cancer prediction. Similarly, the EfficientNet deep learning model effectively extracts handwriting features and identifies patterns associated with Parkinson's disease.

The confusion matrix results indicate that the models correctly classify most of the test samples with minimal misclassification. These results demonstrate that combining machine learning and deep learning techniques can improve the accuracy and reliability of automated disease prediction systems.

Diabetes Prediction Results

Table 2: Performance Metrics for Diabetes Prediction

Metric	Value
Accuracy	0.9714
Precision	0.97
Recall	0.68
F1-Score	0.80
ROC-AUC	0.98

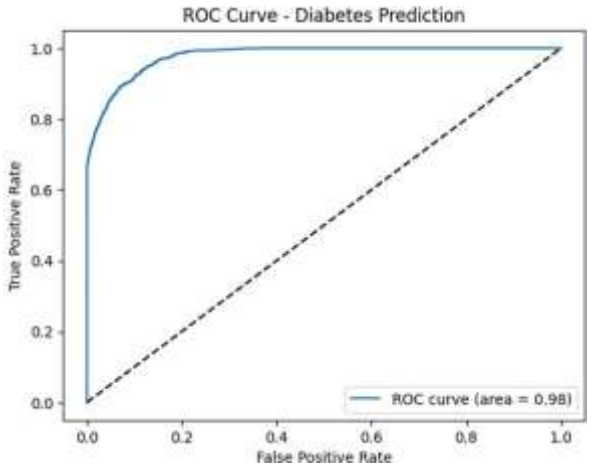

Fig 4: ROC Curve for Diabetes Prediction

The XGBoost model achieved an accuracy of 97.14% for diabetes prediction. The results indicate high precision in identifying diabetic cases, although recall is slightly lower due to dataset imbalance.

Table 3: Performance Metrics for Parkinson's Disease Detection

Metric	Value
Accuracy	0.9638
Precision	0.97
Recall	0.95
F1-Score	0.96
ROC-AUC	0.99

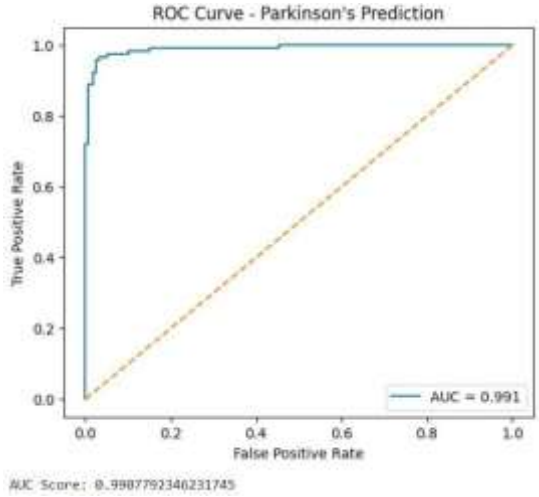

Fig. 5: ROC Curve for Parkinson's Prediction

The EfficientNetB0 deep learning model effectively extracted handwriting features and achieved high classification accuracy in detecting Parkinson's disease.

Table 4: Performance Metrics for Lung Cancer Prediction

Metric	Value
Accuracy	0.977
Precision	0.97
Recall	0.94
F1-Score	0.95
ROC-AUC	0.99

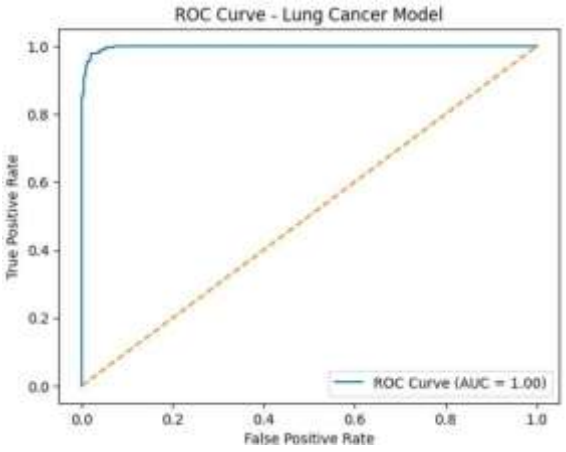

Fig. 6: ROC Curve for Lung Cancer Prediction

The XGBoost classifier demonstrated strong performance in predicting lung cancer risk using medical and lifestyle attributes.

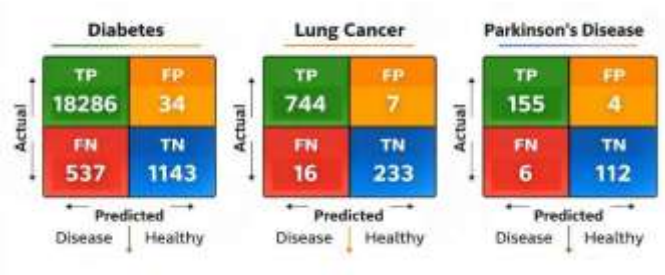

Fig. 7: Confusion Matrix Results for Disease Prediction Models

Table 5: Overall Performance Comparison

Disease	Model	Accuracy
Diabetes	XGBoost	97.14%
Parkinson's	EfficientNetB0	96.38%
Lung Cancer	XGBoost	97.7%

The proposed system achieved an accuracy of 97.14% for diabetes prediction, 96.38% for Parkinson's disease detection, and 97.7% for lung cancer prediction.

7. Gaps Identified in Existing Research

Although significant progress has been made in the field of medical disease prediction using machine learning and deep learning techniques, several limitations still exist in current research. One of the major gaps observed in many existing studies is that most prediction systems are designed to detect only a single disease. These systems are often developed using specific datasets and algorithms that focus on a particular medical condition such as Diabetes, Parkinson's disease, or Lung cancer. As a result, they cannot provide a comprehensive healthcare prediction solution

capable of identifying multiple diseases within a single platform.

Another limitation in existing research is the reliance on limited datasets with restricted patient attributes. Many models are trained on datasets that represent a small or specific population group, which may lead to dataset bias. When such models are applied to new or unseen medical data, their prediction accuracy may decrease due to insufficient diversity in the training data.

Additionally, several traditional prediction systems focus mainly on structured medical datasets and do not effectively utilize image-based data such as handwriting patterns or medical imaging. This restricts their ability to detect neurological conditions or other diseases that require visual analysis.

Another research gap is the lack of user-friendly healthcare platforms. Many proposed models remain at the experimental stage and do not provide interactive systems that allow patients or healthcare professionals to easily access prediction results. Without proper interfaces and guidance mechanisms, these systems cannot be effectively used in real-world healthcare environments.

Furthermore, existing approaches often lack explainability and guidance mechanisms that help users understand prediction results. Many systems simply output classification results without providing meaningful explanations or recommendations.

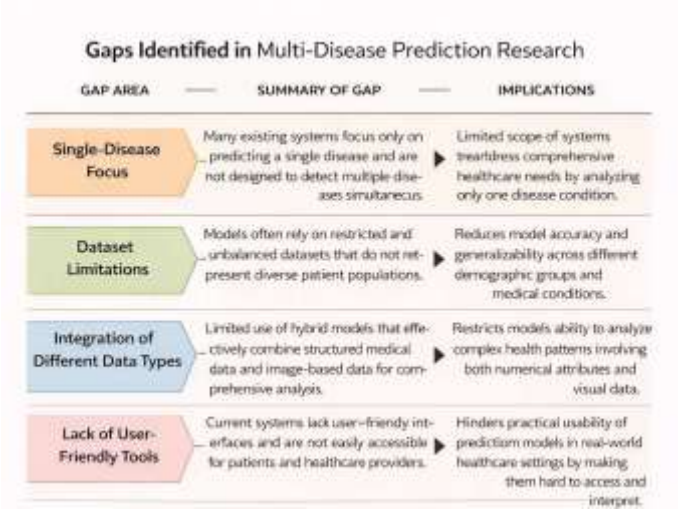

Fig. 8: Gaps Identified in Multi-Disease Prediction Research

To address these research gaps, the proposed system introduces a unified multi-disease prediction framework that integrates machine learning and deep

learning techniques within a single platform.

The system combines structured medical data analysis and image-based prediction models to detect multiple diseases. In addition, an integrated chatbot module provides users with health-related explanations and basic guidance based

on the prediction results, making the system more accessible and practical for real-world healthcare applications.

8. Future Enhancements Suggested in the Literature

Recent research in medical disease prediction highlights several directions for improving the effectiveness and reliability of intelligent healthcare systems. One important enhancement suggested in the literature is the use of larger and more diverse healthcare datasets. Many existing studies rely on limited datasets that may not fully represent different patient populations. Expanding datasets with more medical attributes and patient records can improve the generalization capability of prediction models.

Another potential improvement is the integration of advanced deep learning architectures and hybrid machine learning frameworks. Combining different

algorithms can help capture complex patterns in healthcare data and improve prediction accuracy. Advanced models can also enhance feature extraction from both structured medical data and image-based inputs.

Researchers also suggest improving model robustness and performance by applying advanced techniques such as feature selection, hyperparameter tuning, and data augmentation. These techniques allow prediction models to learn more meaningful patterns and perform better when tested on new datasets.

Another important future direction is the development of explainable artificial intelligence (XAI) techniques for healthcare prediction systems. Explainable models can help users and healthcare professionals understand how prediction results are generated, which improves trust and reliability in AI-based healthcare systems.

Additionally, future systems may integrate real-time healthcare monitoring using wearable devices and Internet of Things (IoT) technologies. Such systems can continuously collect patient health data and provide early disease risk predictions.

Overall, these future enhancements aim to develop more accurate, scalable, and user-friendly healthcare prediction systems that can assist both patients and medical professionals in early disease diagnosis and preventive healthcare.

9. Conclusion

The rapid advancement of artificial intelligence and machine learning technologies has created new opportunities for improving healthcare systems and early disease diagnosis. Predicting diseases at an early stage is important for reducing health risks and improving patient outcomes. In this work, an intelligent multi-disease prediction system was developed to assist in identifying potential health conditions using machine learning and deep learning techniques.

The proposed system integrates different prediction models to detect diseases such as Diabetes, Parkinson's disease, and Lung cancer. Structured healthcare data is analyzed using the XGBoost machine learning algorithm, while handwriting image analysis for Parkinson's disease detection is performed using the EfficientNet deep learning architecture. By combining these models within a unified framework, the system is able to analyze different types of medical data and generate reliable prediction results.

Experimental results demonstrate that the proposed system can effectively identify disease risks with high accuracy and minimal misclassification. The integration of preprocessing techniques and feature extraction methods helps improve model performance and prediction reliability.

In addition, the inclusion of a chatbot module allows users to receive basic health guidance and explanations based on the prediction results, making the system more interactive and user-friendly. Overall, the proposed multi-disease prediction system provides a practical and scalable approach for supporting early disease screening and healthcare decision-making. The

system demonstrates the potential of artificial intelligence to assist healthcare professionals and individuals in identifying possible health risks and taking preventive actions at an early stage.

References

[1]. L. Wang, X. Wang, A. Chen, X. Jin and H. Che, "Prediction of Type-2 Diabetes Risk and Its Effect Evaluation Based on the XGBoost Model," *Healthcare*, vol. 8, no. 3, pp. 247, 2020.
Available: https://doi.org/10.3390/healthcare8030247

[2]. I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas and I. Chouvarda, "Machine Learning and Data Mining Methods in Diabetes Research," *Computational and*

Structural Biotechnology Journal, vol. 15, pp. 104-116, 2017. Available: https://doi.org/10.1016/j.csbj.2016.12.005

[3]. O. Gundogdu, "Efficient Prediction of Early-Stage Diabetes Using XGBoost Classifier with Random Forest Feature Selection Technique," *Multimedia Tools and Applications*, Springer, 2023. Available: https://link.springer.com

[4]. A. Razaq, S. Hussain and M. A. Khan, "A Deep Learning Framework for Early Parkinson's Disease Detection Using Handwriting Patterns," *Diagnostics*,2025. Available: https://www.mdpi.com

[5]. A. Benredjem, M. H. Bencherif and K. Bensmail, "Parkinson's Disease Prediction Using an Attention-Based Multimodal Fusion Framework," *Diagnostics*, 2024.

Available: https://www.mdpi.com

[6]. Y. Lin, J. Chen and H. Liu, "Construction of the XGBoost Model for Early Lung Cancer Prediction Based on Metabolic Indices," *BMC Medical Informatics and Decision Making*, 2023. Available: https://bmcmedinformdecismak.biomedcentral.com

[7]. S. Das, P. Roy and A. Kumar, "Enhanced Deep Learning Approach for Lung Cancer Detection Using CNN Models," *IEEE Access*, 2025. Available: https://ieeexplore.ieee.org

[8]. M. Hossain, S. Rahman and T. Ahmed, "Hybrid CNN-Based Lung Cancer Classification Using Medical Imaging," *IEEE Transactions on Medical Imaging*, 2025. Available: https://ieeexplore.ieee.org

[9]. Z. Rafie et al., "Leveraging XGBoost and Explainable AI for Accurate Prediction of Type-2 Diabetes," *BMC Public Health*, 2025. Available: https://doi.org/10.1186/s12889-025-24953-w

[10]. W. Ji and S. Lin, "The Risk Prediction of Type-2 Diabetes Based on XGBoost," *International Conference on Information Technology*, 2019. Available: https://doi.org/10.23977/meet.2019.93721

[11]. S. Liu, "Diabetes Prediction by KNN, SVM, Random Forest and XGBoost," *Highlights in Science, Engineering and Technology*, 2023. Available: https://doi.org/10.54097/8h8dff76

[12]. K. D. Wardhani and M. Akbar, "Diabetes Risk Prediction Using Extreme Gradient Boosting (XGBoost)," *Jurnal Online Informatika*, 2022. Available: https://doi.org/10.15575/join.v7i2.970

[13]. Z. Chunfu et al., "Diabetes Risk Prediction Based on GA-XGBoost Model," *Computer Engineering Journal*, 2020. Available: https://doi.org/10.19678/j.issn.1000-3428.0054109

[14]. H. Li et al., "Machine Learning-Based Prediction

of Diabetic Patients Using Blood Routine Data," *Methods Journal*, 2024. Available: https://arxiv.org

[15]. F. Ahmed et al., "Automated Early Prediction of Parkinson's Disease Based on Artificial Intelligence Techniques," *Arabian Journal for Science and Engineering*, 2025. Available: https://link.springer.com

[16]. S. Shankar, "CNN-LSTM Hybrid Model for Parkinson's Disease Detection from Handwritten Spirals," *Journal of Image Processing*, 2025. Available: https://imanagerpublications.com

[17]. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016. Available: https://doi.org/10.1145/2939672.2939785

[18]. L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5-32, 2001. Available: https://doi.org/10.1023/A:1010933404324

[19]. S. Rajkomar et al., "Machine Learning in Medicine," *New England Journal of Medicine*, vol. 380, pp. 1347-1358, 2019. Available: https://www.nejm.org

[20]. B. Erickson et al., "Machine Learning for Medical Imaging," *Radiographics*, vol. 37, pp. 505-515, 2017. Available: https://pubs.rsna.org

[21]. J. Esteva et al., "A Guide to Deep Learning in Healthcare," *Nature Medicine*, vol. 25, pp. 24-29, 2019.
Available: https://www.nature.com

[22]. Y. LeCun, Y. Bengio and G. Hinton, "Deep Learning," *Nature*, vol. 521, pp. 436-444, 2015.
Available: https://doi.org/10.1038/nature14539

[23]. A. Krizhevsky, I. Sutskever and G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Neural Information Processing Systems*, 2012.
Available: https://dl.acm.org

[24]. M. Abadi et al., "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems," 2016.
Available: https://tensorflow.org

[25]. F. Chollet, "Deep Learning with Python," *Manning Publications*, 2018.
Available: https://www.manning.com

[26]. S. Farshchiha et al., "Machine Learning Based Methods for Lung Cancer Level Classification," *International Journal of Medical Informatics*, 2025.
Available: https://arxiv.org

[27]. A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau and S. Thrun,
"Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks,"
Nature, vol. 542, no. 7639, pp. 115–118, 2017.
Available: https://doi.org/10.1038/nature21056

[28]. D. Chicco and G. Jurman, "The Advantages of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation,"
BMC Genomics, vol. 21, pp. 6, 2020.
Available: https://doi.org/10.1186/s12864-019-6413-7

[29]. A. Athanasiou et al., "Explainable XGBoost-Based Approach for Cardiovascular Risk in Diabetes Patients," 2020.
Available: https://arxiv.org

[30]. M. Islam et al., "Ensemble Machine Learning Model for Early Diabetes Prediction," 2025.
Available: https://arxiv.org