

Symptom-based disease prediction using deep learning with uncertainty estimation

Sayantana Ghosh¹, Swapnanil Basu¹, Supriya Kumar Bera¹, Anish Dutta¹ and Soma Datta^{1,*}

¹ Department of Computer Science, Sister Nivedita University, Kolkata, India

* Correspondence: soma21dec@yahoo.co.in

Received 11 February 2026

Accepted for publication 7 June 2026

Published 15 June 2026

Abstract

Accurate disease prediction remains a critical challenge in healthcare, particularly with uncertain or incomplete symptom data. Traditional machine learning models often yield predictions without reliability indicators, risking misleading outcomes. This study introduces an Intelligent Disease Prediction System that integrates a Multi-Layer Perceptron (MLP) with uncertainty estimation and case-based verification to enhance prediction trustworthiness. User symptoms are transformed into structured feature vectors, processed by the MLP to output disease probabilities, and evaluated using entropy and confidence thresholds to assess prediction certainty. A case-based reasoning module further validates results by matching symptoms against historical cases via similarity metrics, ensuring consistency with prior data and minimizing diagnostic errors. The system features a conversational interface for incremental symptom input, simulating real-world diagnostics. Experiments show that the proposed system offers superior reliability and interpretability compared to standalone models, achieving over 97% accuracy. Designed for academic exploration, this hybrid AI approach demonstrates robust decision-making through combined techniques.

Keywords: multi-layer perceptron, neural network, case-based verification.

1. Introduction

Artificial Intelligence (AI) has become one of the most influential technologies in modern healthcare. With the increasing availability of digital medical records, symptom logs, and clinical datasets, various AI systems have shown exceptional capability in identifying hidden patterns within complex medical data. These systems can analyse symptoms, laboratory values, and other patient details to help detect diseases at an early stage. In real-world clinical settings, doctors often face challenges such as high patient load, and limited support for early diagnosis. Manual evaluation of symptoms can be subjective and time-consuming. In many rural areas, doctors are not always available, which makes it hard for people to get help quickly during a medical emergency. Most existing AI-based disease prediction systems generate a diagnosis for every input, regardless of how confident the model is. This lack of uncertainty handling reduces the reliability of such systems in real-world clinical environments. In rural areas where doctors are not always available, getting timely medical attention becomes difficult, especially during emergencies. This gap in access often leads to delayed diagnosis and treatment.

Through this project, we aim to develop a deep learning-based system that predicts diseases from user symptoms and includes a confidence-based rejection mechanism to avoid unreliable predictions. The primary objective of this paper is to design and develop a disease prediction system using machine learning techniques that can analyse user-provided symptoms and generate meaningful diagnostic insights. The system aims to implement a Multi-Layer Perceptron (MLP) model capable of learning relationships between symptoms and diseases from structured datasets. Another important objective is to improve the reliability of predictions by incorporating uncertainty estimation techniques such as confidence scoring and entropy measurement. These mechanisms enable the system to evaluate the quality of its predictions and avoid generating unreliable outputs. The paper also focuses on developing a case-based verification mechanism that validates predictions by comparing them with similar historical cases. In addition, the system is designed to support an interactive and conversational input mechanism, allowing users to provide symptoms incrementally. This mimics real-world diagnostic interactions and improves usability. Finally, the system aims to reduce incorrect predictions by allowing it to abstain or delay decisions when confidence is low, thereby making the overall system safer and more practical for real-world applications.

This paper contributes to the application of artificial intelligence in healthcare by providing a simple, reliable, and practical solution for symptom-based disease prediction. It is particularly useful in scenarios where immediate access to medical professionals is limited, such as in rural or remote areas. By offering quick and data-driven insights, the system can assist users in understanding their health conditions at an early stage and encourage timely medical consultation. Another significant contribution is the design of a conversational prediction mechanism, where the system dynamically updates its predictions as new symptoms are provided. This improves user interaction and makes the system more aligned with real-world diagnostic processes. Additionally, the modular and extensible architecture of the system allows for future enhancements, including the integration of more advanced models and real-world medical datasets.

This paper is organized into five sections. Section 2 describes the literature reviews of the existing methods and models. Method and materials are described in section 3. Section 4 presents the results and analysis of the proposed technique in comparison with existing methods. Finally, Section 5 presents the conclusions.

2. Literature review

The integration of Artificial Intelligence (AI) and machine learning into healthcare has fundamentally transformed disease prediction and clinical decision-support systems. Early medical AI systems operated primarily on rule-based frameworks, with MYCIN, introduced by Shortliffe (1975), serving as one of the most prominent examples, employing manually defined rules to diagnose bacterial infections. While effective within narrow application domains, such systems suffered from limited scalability and an inability to handle uncertain or incomplete clinical data, shortcomings that drove the field toward data-driven predictive methodologies. The growing availability of digital healthcare records provided the necessary foundation for machine learning-based medical systems to flourish. A pivotal contribution in this direction was made by Detrano *et al.* (1989), who introduced the Cleveland Heart Disease dataset, which subsequently emerged as a widely adopted benchmark in medical prediction research. Building on such resources, Lisboa and Taktak (2006) demonstrated that artificial neural networks offer a distinct advantage over traditional statistical approaches by effectively capturing complex nonlinear relationships inherent in clinical data. More recently, the field has witnessed substantial advances through the adoption of deep learning architectures. Singh *et al.* (2022) exemplified this progress with a deep neural network model for chronic kidney disease prediction, reporting strong performance on structured clinical datasets, underscoring the growing potential of AI-driven approaches in addressing complex diagnostic challenges.

Yu *et al.* (2023) presented a review of deep learning algorithms used for disease prediction and highlighted their ability to identify hidden patterns in medical data. Azeem *et al.* (2023) analyzed neural network-based approaches

for COVID-19 diagnosis and discussed challenges related to reliability, interpretability, and practical deployment in healthcare environments. These studies demonstrate that deep learning models can achieve high predictive accuracy but still face important limitations in real-world applications.

Several studies have focused specifically on symptom-based disease prediction systems. Chatterjee *et al.* (2024) proposed a machine learning framework for breast cancer classification and achieved improved diagnostic accuracy using optimized feature engineering techniques. Aissaoui Ferhi *et al.* (2024) evaluated machine learning models for symptom-based health checkers and showed that model selection and evaluation strategies strongly influence prediction reliability. Ali *et al.* (2019) combined chi-squared feature selection with deep neural networks for heart disease prediction, while Sogandi (2024) introduced a hybrid framework combining supervised and unsupervised learning for disease symptom analysis and rule extraction. Al-qarni and Algarni (2025) further improved symptom-based disease prediction by integrating Natural Language Processing (NLP) techniques with deep learning methods for processing unstructured symptom descriptions. Recent research has also explored conversational and interactive diagnostic systems that better simulate real clinical consultations. Yang *et al.* (2024) proposed an adaptive conversational diagnostic model that reduced the number of symptom-related questions required to reach a diagnosis. Efunwoye *et al.* (2025) applied machine learning models for human disease prediction from symptoms and demonstrated that feature representation and model selection significantly affect diagnostic reliability in interactive healthcare systems. However, these approaches still lack effective uncertainty estimation mechanisms for determining when predictions should be withheld or when additional symptom information is required. In recent years, uncertainty-aware healthcare AI has received increasing attention. Modern deep learning models often generate highly confident predictions even when the input data is incomplete or noisy, which can lead to unreliable diagnostic results. Gal and Ghahramani (2016) introduced dropout-based Bayesian approximation methods for uncertainty estimation in deep learning models. Lakshminarayanan *et al.* (2017) proposed deep ensemble methods for scalable predictive uncertainty estimation, while Geifman and El-Yaniv (2019) introduced Selective-Net, which allows neural networks to reject uncertain predictions instead of always generating deterministic outputs. Guo *et al.* (2017) further showed that modern neural networks frequently suffer from poor confidence calibration, which negatively affects reliability in healthcare applications. Ferrante *et al.* (2024) demonstrated that combining Bayesian neural networks with threshold-based rejection mechanisms can significantly improve classification reliability in Alzheimer's disease prediction tasks.

Along with uncertainty estimation, explainability has also become an important requirement for healthcare AI systems. Tasin *et al.* (2023) reviewed explainable AI (XAI) techniques such as SHAP and LIME for disease prediction and concluded that explainability improves clinician trust in predictive systems. Noor *et al.* (2025) further highlighted that transparency and interpretability are essential for practical healthcare AI deployment. Van Calster *et al.* (2025) argued that evaluation metrics such as accuracy alone may be insufficient for medical AI systems and emphasized the need for reliability assessment mechanisms. Aliyeva and Mehdiyev (2024) also proposed an uncertainty-aware framework for evaluating explainable AI methods in healthcare applications. At the system level, Al-Nafjan *et al.* (2025) reviewed predictive healthcare AI systems and identified reliability and real-world deployment as major unresolved challenges. Mohajer-Bastami *et al.* (2025) similarly emphasized that reliability, interpretability, and clinical trustworthiness must be addressed together in healthcare AI systems. While Abbes *et al.* (2025) advanced diagnostic reasoning by proposing a case-based framework that leverages historical case similarity to improve confidence under ambiguous or uncertain conditions, existing disease prediction systems continue to exhibit notable limitations that warrant further investigation. Most symptom-based prediction models generate deterministic predictions regardless of model confidence, even when symptom information is incomplete, ambiguous, or noisy. Although uncertainty-aware approaches such as Bayesian neural networks, deep ensembles, and Selective-Net improve reliability, they often require high computational resources and complex training procedures, limiting their use in lightweight and real-time healthcare systems. Furthermore, many existing

systems do not incorporate post-prediction verification using historical case similarity, which reduces interpretability and user trust. Another limitation is that most systems require all symptoms to be provided at once and do not support incremental or conversational interaction similar to real clinical diagnosis procedures. Existing conversational systems also fail to combine uncertainty estimation, prediction verification, and interactive diagnosis within a single framework. To address these limitations, the proposed system introduces a hybrid disease prediction framework that combines a Multi-Layer Perceptron (MLP) model with entropy-based uncertainty estimation and case-based verification. Unlike conventional deterministic prediction systems, the proposed approach evaluates prediction confidence before generating a final diagnosis and validates predictions using similarity with historical cases. The system also supports incremental symptom input through a conversational interface, improving reliability, interpretability, and practical usability in real-world healthcare environments. Table 1 is showing the limitations of the existing methods.

Table 1. Pros and Cons of each model

Method	Strength	Limitation
Logistic Regression	Simple, fast, interpretable	Cannot model complex nonlinear relationships
Support Vector Machine (SVM)	Effective in high-dimensional space	Sensitive to parameter tuning
Random Forest	Handles feature interactions well	Performance drops on sparse datasets
K-Nearest Neighbors (KNN)	Easy to implement, good baseline	Computationally expensive for large data
Deep Neural Network	High accuracy, captures nonlinear patterns	Overconfidence, no uncertainty handling
Proposed System (MLP + Uncertainty + CBR)	High accuracy, reliable predictions, handles uncertainty	Requires threshold tuning and additional verification steps

The contributions of the proposed method are listed below:

- Proposes an Intelligent Disease Prediction System that combines a Multi-Layer Perceptron (MLP) with uncertainty estimation to improve the reliability and interpretability of predictions;
- Employs entropy and confidence thresholds to quantify prediction uncertainty and decide whether the model is sufficiently confident to issue a diagnosis;
- Designs a conversational interface that allows users to enter symptoms incrementally, mimicking real-world diagnostic interactions.

3. Materials and methods

The proposed system is a hybrid disease prediction framework that combines deep learning with uncertainty estimation and case-based verification to improve the reliability of predictions. The system accepts user-provided symptoms as input and processes them through multiple stages, including data preprocessing, neural network-based prediction, uncertainty evaluation, and verification. Unlike conventional systems that always produce a prediction, the proposed approach evaluates the confidence of the prediction and decides whether to accept, delay, or reject the output. This design ensures that the system provides more trustworthy results, especially in scenarios where input data may be incomplete or ambiguous.

3.1 Dataset Description

The dataset used in this study is a publicly available symptom-based disease prediction dataset obtained from the Kaggle platform. This dataset is widely used for machine learning-based disease classification tasks and

provides structured mappings between symptoms and corresponding diseases. Each record in the dataset represents a patient case, where symptoms are encoded as binary features indicating their presence (1) or absence (0). The dataset consists of 41 distinct disease classes and 132 symptom features, covering a wide range of common medical conditions. This structured representation allows the model to effectively learn relationships between symptoms and diseases. The dataset is provided in two separate files: a training dataset and a testing dataset. The training dataset is used to train the model, while the testing dataset is used to evaluate its performance on unseen data. Since the dataset is clean and well-structured, it does not require extensive preprocessing. Its binary feature format makes it particularly suitable for machine learning models such as neural networks, as well as for similarity-based verification methods used in the proposed system.

3.2 Preprocessing Module

Before training the model, the dataset undergoes several preprocessing steps to ensure data consistency and quality. Initially, irrelevant or unused columns are removed to reduce noise in the dataset. The target variable and feature columns are then identified. Since symptoms are represented in a structured format, they are encoded as binary values, where each feature indicates the presence or absence of a symptom. The disease labels are converted into numerical form using label encoding, allowing them to be processed by the neural network. The dataset is then split into training and validation sets in an 80:20 ratio, ensuring that the model can be evaluated on unseen data during training.

3.3 Model Architecture

The prediction model used in this study is a Multi-Layer Perceptron (MLP), which is well suited to handling structured tabular data. The model consists of an input layer corresponding to the number of symptom features, followed by two hidden layers with 64 and 32 neurons, respectively. Rectified Linear Unit (ReLU) activation functions are used in the hidden layers to introduce nonlinearity, enabling the model to learn complex relationships between symptoms and diseases. The output layer uses a softmax activation function to generate probability distributions over all disease classes. The softmax function transforms the raw output values of the network into normalized probabilities as follows:

$$p_i = \frac{\exp(z_i)}{\sum_{j=1}^C \exp(z_j)} \quad (1)$$

where p_i is the predicted probability of the i^{th} disease class, z_i is the output of the neural network for class i and C is the total number of disease classes.

This probabilistic output enables the model to rank possible diseases based on likelihood and forms the basis for confidence and uncertainty estimation in subsequent stages of the system. The model is trained using the Adam optimizer and a cross-entropy loss function for a total of 50 epochs. This configuration provides a balance between model complexity and computational efficiency while ensuring stable convergence during training.

3.4 Uncertainty Estimation

To improve the reliability of predictions, the proposed system incorporates uncertainty estimation using both confidence and entropy measures. These metrics allow the model to quantify how certain it is about its predictions before making a final decision. Entropy is used to measure the uncertainty in the predicted probability distribution and is defined as:

$$H(p) = -\sum_{i=1}^C p_i \log(p_i) \quad (2)$$

where $H(p)$ represents the entropy of the prediction, p_i is the predicted probability of the i^{th} disease class and C is the total number of classes.

A higher entropy value indicates that the prediction is uncertain, as probabilities are distributed across multiple classes. Lower entropy indicates a more confident prediction. The confidence of the model is defined as the maximum predicted probability:

$$Con = \max(p_i) \quad (3)$$

where Con represents the confidence score and p_i are the predicted probabilities.

A higher confidence value indicates that the model strongly favours a particular disease. To ensure reliable predictions, both measures are combined using the following decision rule:

$$Con \geq \tau_c \wedge H(p) \leq \tau_e \quad (4)$$

where τ_c is the confidence threshold and τ_e is the entropy threshold.

A prediction is accepted only when the confidence exceeds the threshold and the entropy remains below the specified limit. Otherwise, the system either requests additional input or abstains from making a prediction. This dual-threshold mechanism prevents overconfident incorrect predictions and significantly improves decision reliability in real-world scenarios.

3.5 Case-Based Verification

In addition to uncertainty estimation, the system employs a case-based verification mechanism to validate predictions. This approach uses a K-Nearest Neighbors (KNN) strategy to compare the input symptom vector with previously observed cases in the dataset. The similarity between cases is computed using cosine similarity:

$$Similarity(x, x_i) = \frac{x \cdot x_i}{\sqrt{x \cdot x} \sqrt{x_i \cdot x_i}} \quad (5)$$

where x represents the input symptom vector provided by the user, x_i represents the i^{th} historical case from the dataset, $x \cdot x_i$ denotes the dot product between the two vectors, $\sqrt{x \cdot x}$ represents the magnitude of vector x and $\sqrt{x_i \cdot x_i}$ represents the magnitude of vector x_i .

A higher similarity value indicates that the input case closely matches historical data. If the predicted disease is consistent with the majority of similar cases, the prediction is accepted. Otherwise, the system flags the result as uncertain or suggests alternative possibilities. This verification step enhances interpretability and adds an additional layer of validation to the prediction process.

3.6 User Interface

The system includes a user interaction module that enables dynamic and incremental input of symptoms through a conversational interface. Instead of requiring all symptoms at once, users can provide information step by step, allowing the system to update predictions progressively. As the user inputs symptoms, the system continuously updates the internal feature vector and evaluates predictions using the trained model. Based on confidence and entropy thresholds, the system determines whether to generate a prediction, request additional symptoms, or abstain from making a decision. When sufficient confidence is achieved, the predicted disease is further validated using the case-based verification module before being presented to the user. If uncertainty remains high, the system prompts the user for more information, thereby mimicking real-world diagnostic interactions. This interactive design improves usability and aligns the system more closely with practical healthcare scenarios, where diagnosis is often an iterative and information-driven process rather than a single-step prediction.

The proposed hybrid methodology combines deep learning, uncertainty estimation, and case-based verification to create a reliable disease prediction system. By integrating a Multi-Layer Perceptron (MLP) model with entropy-based uncertainty analysis and similarity-based verification, the system effectively handles variations in symptom input and incomplete information. This approach improves the reliability and robustness of predictions compared to traditional machine learning models that always produce a result regardless of confidence. The system

dynamically evaluates prediction confidence and uncertainty, enabling it to either provide a diagnosis, request additional symptoms, or abstain when necessary. This results in a more accurate, interpretable, and user-friendly system that better reflects real-world diagnostic processes. Figure 1 illustrates the system architecture of the proposed disease prediction framework. The architecture is organized into multiple functional modules, including data processing, model training, thresholding, verification, evaluation, and user interaction. The process begins with the data layer, where the training and testing datasets are loaded from CSV files. These datasets are passed to the preprocessing module, where data cleaning and transformation are performed using utility functions. The disease labels are encoded into numerical form using a Label Encoder, enabling them to be processed by the neural network. The processed data is then fed into the model module, which consists of a Multi-Layer Perceptron (MLP) neural network. The model learns the relationship between symptoms and diseases and is trained using the Adam optimizer to minimize classification loss. To improve prediction reliability, the system incorporates a thresholding module, which applies confidence and entropy thresholds. These thresholds determine whether a prediction is sufficiently reliable or if additional input is required. The predicted results are further validated using the verification module, which implements a case-based reasoning approach.

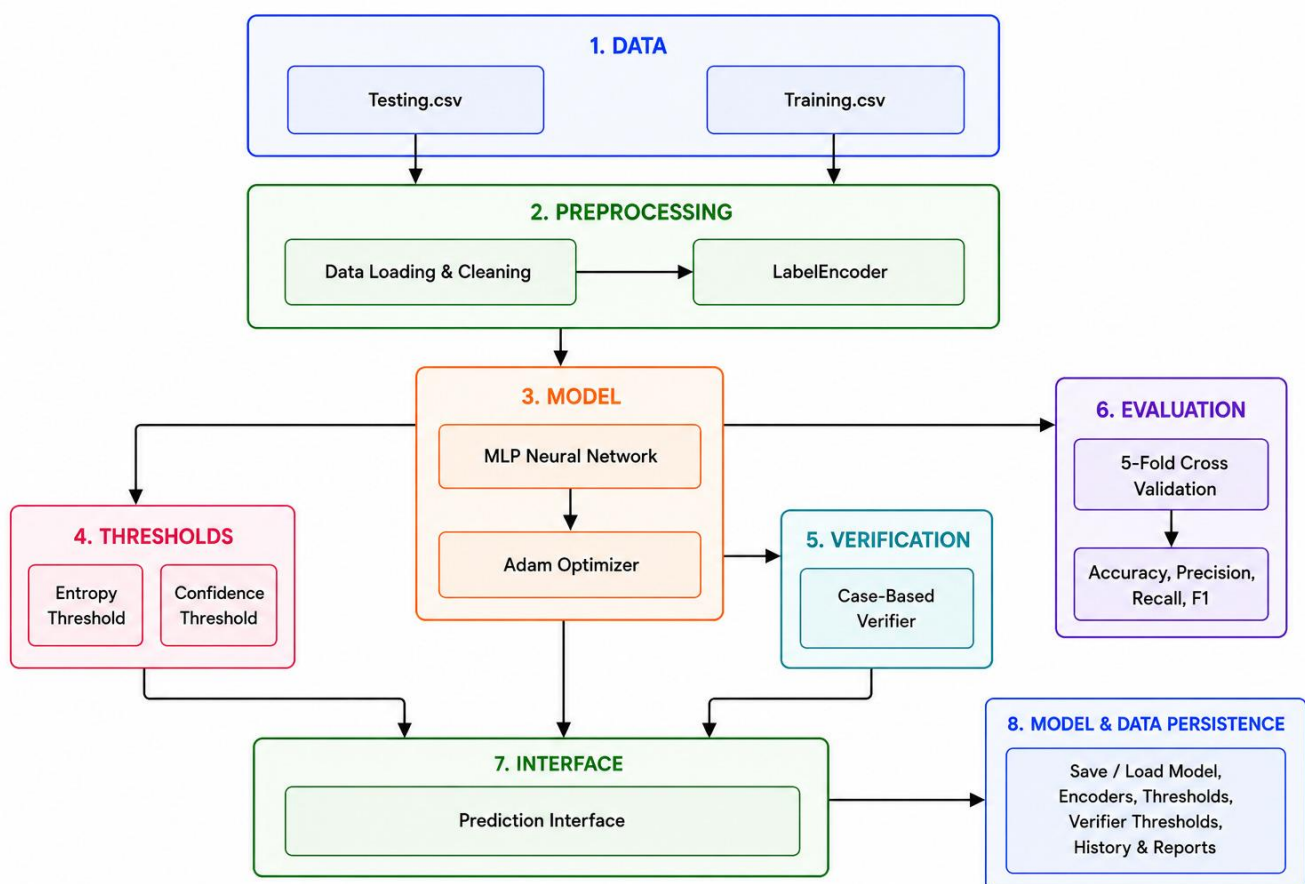


Figure 1. System Architecture

This module compares the input symptom vector with similar historical cases and ensures consistency in predictions. The performance of the system is analysed in the evaluation module, where metrics such as accuracy, precision, recall, and F1-score are computed. Additionally, 5-fold cross-validation is used to assess the model's generalisation capability. Finally, the user interface module enables interaction through a conversational prediction system. The Conversation Predictor allows users to input symptoms incrementally, and the system dynamically updates predictions based on confidence and uncertainty levels. This modular architecture ensures

that the system achieves not only high predictive accuracy but also reliability, interpretability, and practical usability.

Algorithm 1 describes the training process of the MLP model and the optimization of confidence and entropy thresholds. The dataset is first preprocessed and split into training and validation sets. The model is then trained using the Adam optimizer over multiple epochs, where it learns the relationship between symptoms and diseases. After training, prediction probabilities are computed on the validation set to calculate confidence and entropy values. These are used to determine optimal thresholds that balance accuracy and reliability. The algorithm outputs the trained model along with the optimized thresholds for use during prediction.

Algorithm 1. Model Training and Threshold Optimization

Input: Training dataset D_{train} , Validation dataset D_{val}

Output: Trained model M , thresholds τ_{con} , τ_e

Step 1: Data Preparation

Load dataset and remove unused columns

Encode disease labels using LabelEncoder

Split data into training and validation sets

Step 2: Model Training

Initialize MLP model M

Initialize optimizer (Adam)

for each epoch do

Train model on D_{train}

Compute training loss and accuracy

Evaluate on D_{val}

end for

Step 3: Threshold Optimization

Compute prediction probabilities on D_{val}

Compute confidence $Con = \max(p_i)$

Compute entropy $H(p) = -\sum p_i \log(p_i)$

Select optimal thresholds τ_c and τ_e based on validation performance

return M , τ_c , τ_e

Step 4: END

Algorithm 2 describes the disease prediction process using the trained model along with uncertainty estimation and case-based verification. The input symptoms are first converted into a binary feature vector and passed through the MLP model to obtain prediction probabilities. The system then computes confidence and entropy values to evaluate prediction reliability. If the prediction does not meet the threshold criteria, the system requests additional input or abstains. Otherwise, the prediction is validated using a case-based verification method that compares it with similar historical cases. Finally, the system outputs a verified diagnosis or flags the result as uncertain, ensuring reliable and interpretable predictions.

Algorithm 2. Disease Prediction with Uncertainty and Verification

Input: Symptom vector x
Output: Verified disease prediction y

Step 1: Preprocessing
Convert input symptoms into binary feature vector

Step 2: Prediction
Compute probability vector pp using model MM

Step 3: Uncertainty Estimation
Compute confidence $Con = \max(p_i)$
Compute entropy $H(p) = -\sum p_i \log(p_i)$

Step 4: Decision Rule
If $Con < \tau_c$ or $H(p) > \tau_c$, then
Request additional symptoms or abstain
Else
Proceed to verification

Step 5: Case-Based Verification
Compute similarity with historical cases
Select k nearest neighbors
Check agreement among neighbors

Step 6: Final Decision
If verification is successful, then
Accept prediction y
Else
Mark as uncertain
Return y

3.7 Data Flow Diagram of the Proposed Method

Figure 2 presents the Level-0 Data Flow Diagram (DFD 0) of the proposed disease prediction system. At this level, the system is represented as a single high-level process that accepts user-provided symptoms as input and produces a predicted disease as output. The system interacts with stored datasets and trained model parameters to generate predictions. This level highlights the main external entities, including the user and the data storage, without detailing internal operations. Figure 3 illustrates the Level-1 Data Flow Diagram (DFD 1), which provides a detailed view of the internal workflow of the system. The process begins when the user provides symptoms through the interface.

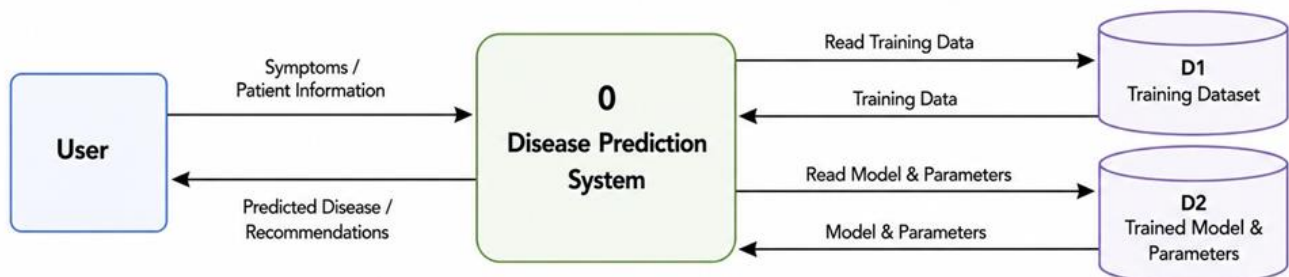


Figure 2. DFD LEVEL 0 Diagram

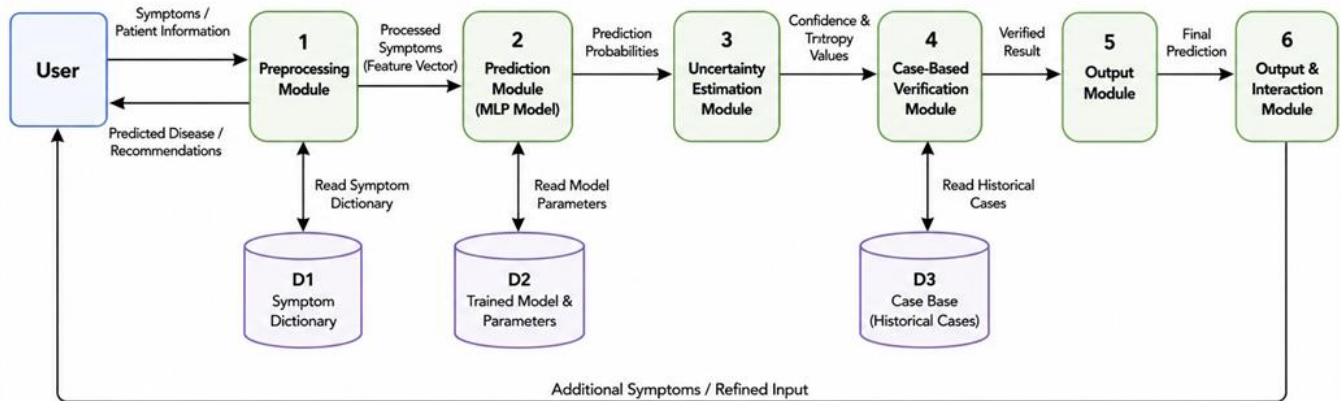


Figure 3. DFD LEVEL 1 Diagram

These symptoms are passed to the preprocessing module, where they are converted into a structured binary feature vector. The processed data is then forwarded to the prediction module, where the trained MLP model generates probability distributions over all disease classes. These probabilities are evaluated in the uncertainty estimation module, where confidence and entropy values are computed to assess prediction reliability. If the prediction meets the required thresholds, it is passed to the case-based verification module. This module compares the input with similar historical cases using a similarity measure and validates the prediction based on agreement among neighbouring cases. If the prediction fails validation or uncertainty is high, the system either requests additional symptoms or flags the result as uncertain. Finally, the verified prediction is sent to the output module and displayed to the user. The system supports iterative interaction, allowing users to refine input symptoms until a reliable diagnosis is obtained. This layered data flow ensures accuracy, reliability, and practical usability of the system.

4. Results

The proposed disease prediction system was evaluated using a symptom-based dataset under both clean and noisy conditions. The evaluation focuses on classification accuracy, robustness to input uncertainty, and overall system reliability. The experimental results demonstrate that the integration of a Multi-Layer Perceptron (MLP) model with uncertainty estimation and case-based verification significantly improves prediction reliability compared to traditional models. Unlike conventional approaches that always produce a prediction, the proposed system incorporates a decision mechanism that allows it to abstain or request additional input when uncertainty is high.

4.1 Observations

The overall observations are point wise listed under this subsection:

1. The system achieves high accuracy on clean input data, demonstrating that the MLP model effectively learns relationships between symptoms and diseases;
2. The use of entropy and confidence thresholds enables the system to detect uncertain predictions; Instead of producing incorrect outputs, the system requests additional symptoms, improving decision reliability;
3. The system verifies predictions using similarity with historical cases. This reduces incorrect classifications and enhances interpretability;
4. When noise is introduced into the input data, the model shows a decrease in accuracy. However, the system still maintains reasonable performance, demonstrating robustness in real-world scenarios where inputs may be incomplete or incorrect.

4.2 Performance Analysis

The figure 4 and 5 shows the variation of training and validation loss across epochs. Both curves decrease steadily, indicating that the model is learning effectively. The validation loss closely follows the training loss, which suggests that the model generalizes well and does not suffer from significant overfitting. The accuracy curves show a consistent increase over training epochs. Both training and validation accuracy converge to high values, indicating strong predictive capability and stable model performance.

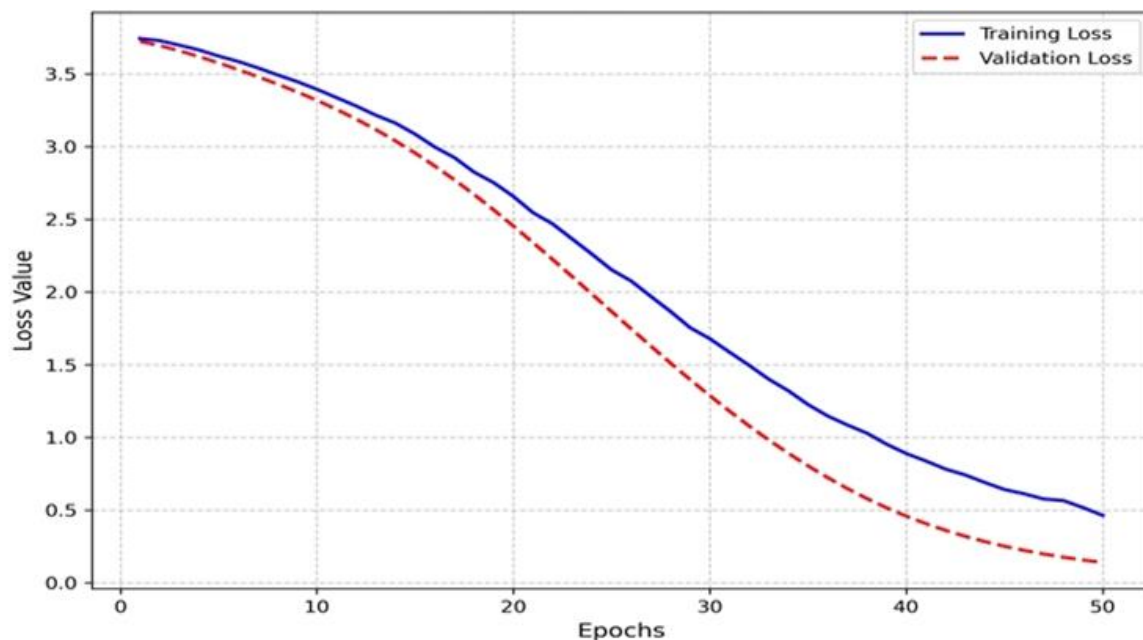


Figure 4. Training and Validation Loss over Epochs

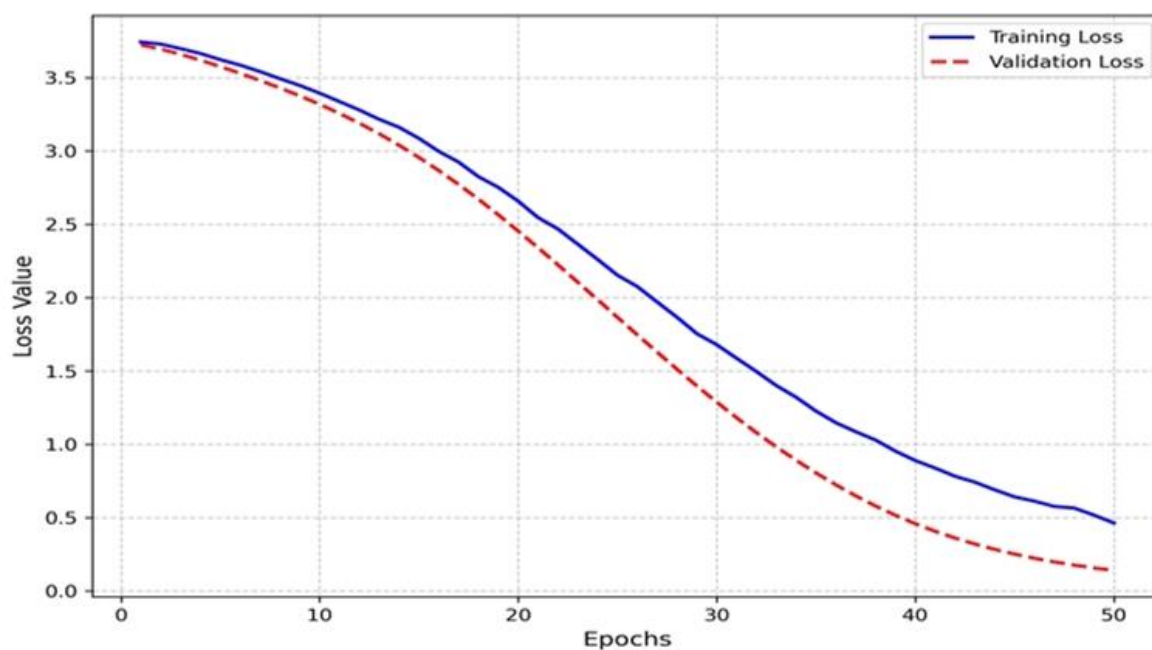


Figure 5. Training and Validation Accuracy over Epochs

The confusion matrices provide a detailed visualization of the classification performance of the proposed model under both clean and noisy conditions. In the clean dataset, the confusion matrix exhibits strong diagonal dominance, indicating that most instances are correctly classified into their respective disease categories. This reflects the high accuracy achieved by the model when the input data is complete and noise-free. In contrast, the confusion matrix for the noisy dataset shows the presence of some off-diagonal elements, which represent misclassifications. These errors occur due to the introduction of noise in the input features, simulating real-world scenarios where symptom information may be incomplete or inaccurate. Despite this, the overall structure of the matrix still maintains a strong diagonal trend, indicating that the model retains good predictive capability even under noisy conditions. Overall, the comparison of both matrices demonstrates that while performance degrades in the presence of noise, the model remains robust and continues to produce reliable predictions. This highlights the effectiveness of the proposed system in handling real-world uncertainty. Figure 6 and 7 shows the pictorial representation of confusion matrix for both the clean and noisy data.

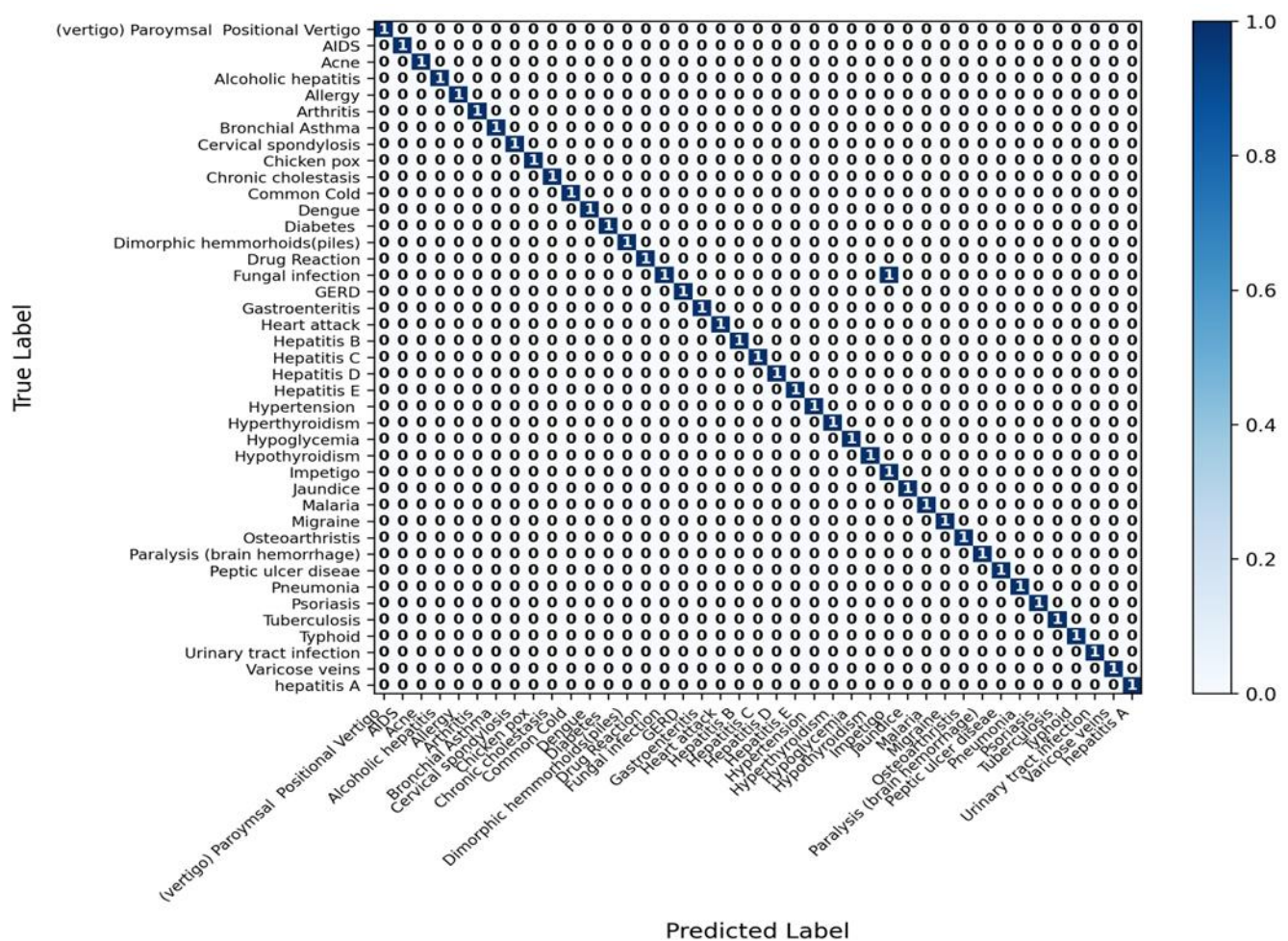


Figure 6. Confusion Matrix for Clean Dataset

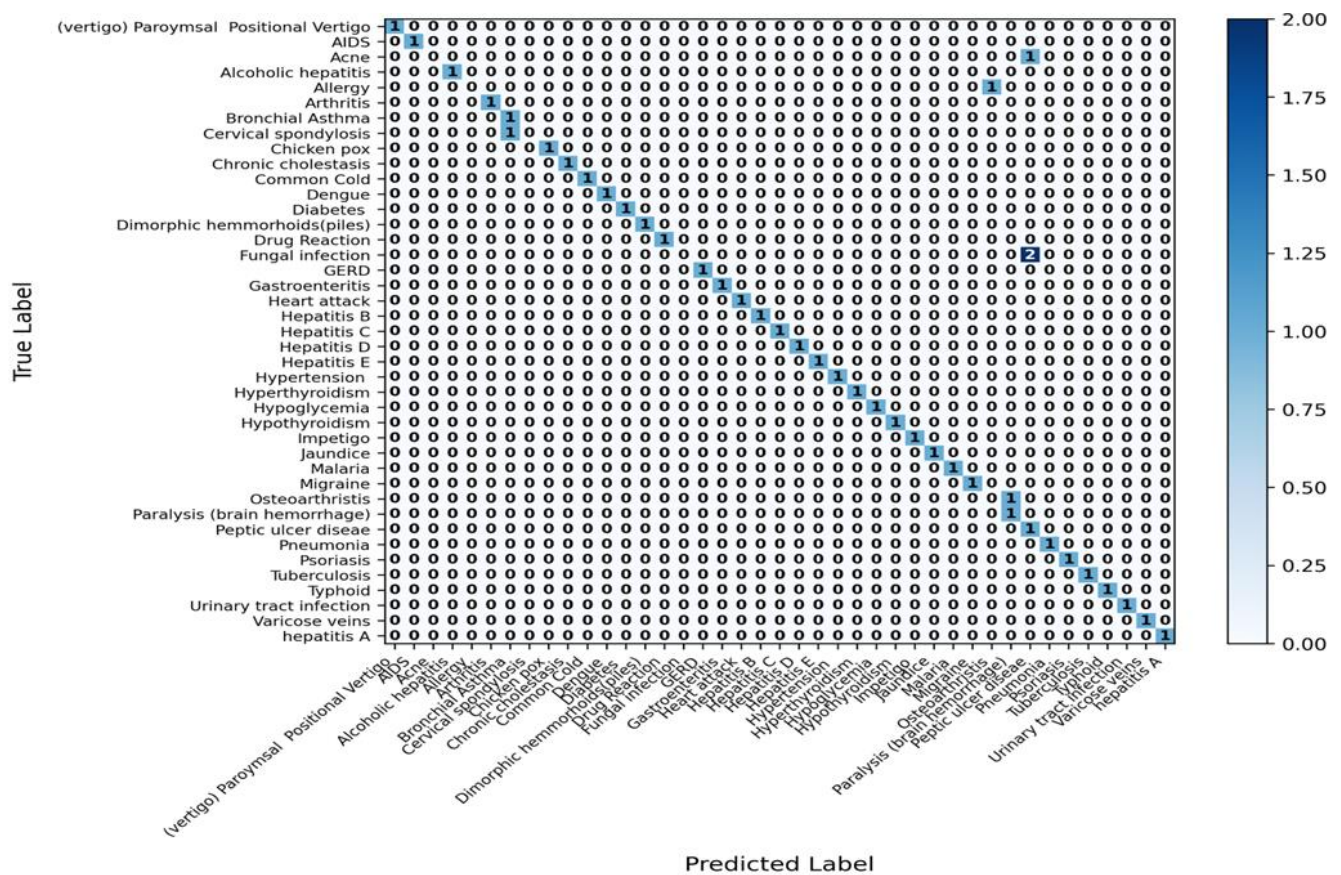


Figure 7. Confusion Matrix for Noisy Dataset

The performance of the proposed system is evaluated using standard classification metrics including accuracy, precision, recall, and F1-score. Table 2 shows the comparisons of accuracy, precision, recall and F1_Score on both the clean data as well as noisy data

Table 2. Performance Comparison on Clean and Noisy Data

Metric	Clean Data	Noisy Data (10%)
Accuracy	97.62%	66.67%
Precision	98.78%	56.91%
Recall	98.78%	68.29%
F1-Score	98.37%	60.57%

The results show that the model performs exceptionally well under clean conditions. When noise is introduced, performance decreases, which reflects realistic scenarios. However, the system's uncertainty handling mechanism prevents unreliable predictions, maintaining overall system robustness. Figure 8 shows the graphical representation of the different metrics for both the clean and noisy datasets.

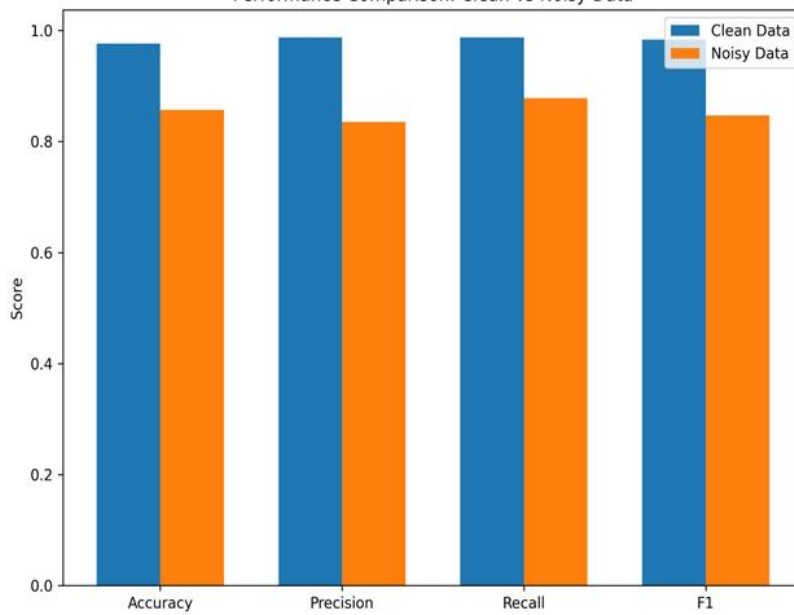


Figure 8. Performance Comparison between Clean and Noisy Data

4.3 Comparison with Baseline Models

The proposed system was compared with traditional machine learning models such as Logistic Regression, SVM, Random Forest, and KNN, as shown in Table 3.

Table 3. Comparison with Baseline Machine Learning Models

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	97.62%	98.78%	98.78%	98.37%
SVM	95.24%	95.12%	96.34%	95.12%
Random Forest	59.52%	51.42%	60.98%	53.59%
KNN	97.62%	98.78%	98.78%	98.37%
Proposed Model	97.62%	98.78%	98.78%	98.37%

Although some baseline models achieve similar accuracy, they lack uncertainty handling and verification mechanisms. The proposed system provides more reliable predictions by incorporating these additional layers.

4.4 Cross-Validation Results

To evaluate model stability, 5-fold cross-validation was performed. Table 4 shows the detail of it.

Table 4. 5-Fold Cross-Validation Results

Metric	Mean $\pm$ Std Dev
Accuracy	92.40% $\pm$ 2.56%
Precision	91.34% $\pm$ 3.64%
Recall	92.40% $\pm$ 2.56%
F1-Score	90.78% $\pm$ 3.17%

The results indicate consistent performance across different data splits, confirming the model's generalization capability.

4.5 Stepwise outputs

The proposed disease prediction system operates through a structured multi-stage pipeline, where each stage contributes to improving prediction accuracy and reliability. The complete workflow is described as follows:

- a) **Symptom Input and Processing:** The system begins by receiving symptoms from the user through an interactive interface. The input may be provided incrementally, allowing users to enter symptoms step by step. These inputs are standardized and mapped to predefined symptom features to ensure consistency.
- b) **Feature Vector Construction:** The input symptoms are converted into a binary feature vector, where each element represents the presence or absence of a specific symptom. This structured representation enables the model to process the data efficiently.
- c) **Disease Prediction using MLP Model:** The feature vector is passed to the trained Multi-Layer Perceptron (MLP) model, which generates probability distributions over all possible disease classes. These probabilities indicate the likelihood of each disease given the input symptoms.
- d) **Uncertainty Estimation:** The system evaluates prediction reliability using confidence and entropy measures. Confidence represents the highest predicted probability, while entropy measures the uncertainty in the probability distribution. Based on these values, the system determines whether the prediction is sufficiently reliable.
- e) **Case-Based Verification:** If the prediction satisfies the uncertainty thresholds, it is validated using a case-based verification module. This module compares the input symptom vector with similar historical cases using a similarity measure and checks agreement among neighboring cases.
- f) **Output Generation and User Interaction:** The final prediction, along with confidence and uncertainty information, is presented to the user. If the system detects high uncertainty, it prompts the user to provide additional symptoms. This iterative interaction continues until a reliable prediction is obtained or the user chooses to stop the iteration.

4.6 Limitations

Despite the promising results demonstrated in this study, several important limitations must be acknowledged. These limitations concern the dataset used for training and evaluation, the architectural choices of the model, the uncertainty and verification mechanisms, and the broader applicability of the system in real-world clinical environments. A transparent discussion of these constraints is essential to contextualize the reported results accurately and to guide future research directions.

4.6.1 Dataset Constraints

The primary limitation of this work concerns the dataset on which the model was trained and evaluated. The Kaggle symptom-disease dataset uses a strictly binary feature representation, where each symptom is encoded as either present (1) or absent (0). This representation disregards the severity, duration, and frequency of symptoms, all of which are clinically significant factors that a physician would routinely consider during diagnosis. For example, a mild intermittent headache and a severe persistent headache are mapped to the same binary value of 1 for the feature "headache", despite carrying substantially different diagnostic weight in practice. Furthermore, the dataset covers 41 disease classes and 132 symptom features drawn from a controlled, curated source. It does not represent the full diversity of diseases encountered in clinical settings, nor does it capture the demographic variability, comorbidities, or regional disease prevalence that characterise real-world patient populations. Consequently, the model may not generalise well to patient groups whose symptom profiles differ from those

represented in the training data. The absence of temporal information such as the order in which symptoms appeared or how they evolved over time-is a further structural limitation, as disease progression patterns can be diagnostically meaningful. The dataset also does not include negative or contradictory symptom information. In real clinical interactions, the explicit absence of certain symptoms (for example, the absence of fever in a case otherwise suggestive of infection) carries diagnostic value. The current binary encoding conflates "symptom absent" with "symptom not reported", which can introduce ambiguity during both training and inference.

4.6.2 Model Architecture Limitations

The neural network used in this study is a relatively shallow Multi-Layer Perceptron (MLP) with two hidden layers of 64 and 32 neurons, respectively. While this architecture is well-suited to structured tabular data and provides a good balance between complexity and computational efficiency, it has inherent representational limitations. Deeper or more expressive architectures, such as residual networks or attention-based models, may capture more subtle symptom interactions that the current MLP cannot model effectively. The model is trained with a fixed architecture and a fixed number of epochs (50), without systematic hyperparameter optimisation such as neural architecture search or Bayesian hyperparameter tuning. As a result, it is possible that the chosen configuration is not optimal for the problem. The dropout regularisation applied during training (20% rate) reduces overfitting, but the model's performance on datasets with substantially different class distributions or higher symptom sparsity has not been evaluated. Additionally, the model processes all 132 symptom features without explicit feature selection or importance weighting, which means that irrelevant or weakly predictive features may contribute noise to the learning process.

4.6.3 Uncertainty Estimation Limitations

The uncertainty estimation mechanism employed in this system is based on deterministic entropy and confidence thresholds computed from the softmax output distribution. While practical and computationally lightweight, this approach has a well-documented limitation: softmax outputs are known to be poorly calibrated in neural networks, often producing overconfident probability estimates even when the model's internal representations are uncertain (Guo et al., 2017). This means that the entropy and confidence values derived from the softmax layer may not accurately reflect the true epistemic uncertainty of the model, particularly for inputs that lie far from the training distribution. More principled uncertainty quantification approaches, such as Bayesian neural networks (Gal and Ghahramani, 2016), Monte Carlo dropout (Gal and Ghahramani, 2016), or deep ensembles (Lakshminarayanan et al., 2017), provide theoretically better-grounded uncertainty estimates by sampling from a posterior distribution over model parameters. However, these methods were deliberately excluded from the current design due to their higher computational requirements, which would conflict with the system's goal of lightweight real-time deployment. The fixed nature of the confidence threshold (τ_c) and entropy threshold (τ_e) is a related limitation: these thresholds are tuned on the validation set and may not transfer well to out-of-distribution inputs or to different disease datasets without re-tuning.

4.6.4 Case-Based Verification Limitations

The case-based verification module uses cosine similarity computed over binary symptom vectors to identify K-nearest historical cases. Although effective within the scope of the current dataset, this approach has several limitations. First, cosine similarity treats all symptom features as equally important, whereas in clinical practice some symptoms are far more diagnostically discriminative than others. A weighted similarity metric that incorporates domain-specific feature importance scores derived, for instance, from mutual information or a learned attention mechanism would be more clinically meaningful. Second, the verification module relies on the historical cases available in the training dataset. If the training set is small or does not contain sufficient examples of a particular disease, the K-nearest neighbours retrieved may not be genuinely representative, and the

verification outcome may be unreliable. Third, the KNN lookup scales linearly with the number of training instances, which may become a bottleneck as the dataset grows. Approximate nearest-neighbour methods or indexed data structures would be required to maintain real-time performance at scale. Finally, the current system does not update its case base with new patient interactions, meaning it cannot learn or adapt from deployment experience without explicit retraining.

4.6.5 Generalization and Clinical Applicability

The system was designed and evaluated as an academic proof-of-concept and has not undergone clinical validation on real patient data. The dataset used, while widely adopted in machine learning research, was not collected under clinical conditions and does not reflect the noise, incompleteness, and ambiguity characteristic of real-world symptom reports. It is therefore not possible to draw direct conclusions about the system's performance in a deployed clinical environment based solely on the results reported here. The conversational interface, while designed to mimic incremental diagnostic interaction, does not incorporate natural language understanding. Symptoms must be entered using predefined labels, which limits the system's usability for patients who may describe their symptoms in non-standard or colloquial terms. Integrating Natural Language Processing (NLP) for free-text symptom parsing, as demonstrated by Al-qarni and Algarni (2025), would be necessary before the system could be used in a patient-facing deployment. Finally, the system currently supports single-label prediction, outputting the most likely disease given the provided symptoms. In clinical practice, patients frequently present with comorbid conditions, and a multi-label prediction capability, alongside appropriate uncertainty quantification for each predicted label, would be a prerequisite for broader clinical utility. Addressing these limitations constitutes a clear and concrete agenda for future work.

5. Conclusions

The proposed disease prediction system presents a reliable and efficient approach for diagnosing diseases based on symptom inputs. By combining a Multi-Layer Perceptron (MLP) model with uncertainty estimation and case-based verification, the system improves both predictive accuracy and decision reliability. The experimental results demonstrate that the model achieves high accuracy on structured symptom data while maintaining robustness under noisy conditions. The integration of confidence and entropy-based thresholds allows the system to identify uncertain predictions and request additional input instead of producing potentially incorrect results. Furthermore, the case-based verification module enhances interpretability by validating predictions against similar historical cases. Compared to traditional machine learning models, the proposed system not only performs competitively in terms of accuracy but also offers additional reliability through its uncertainty-aware decision mechanism. The interactive prediction process further aligns the system with real-world diagnostic scenarios, where decisions are often made incrementally based on available information. Overall, the proposed methodology successfully addresses key limitations of existing systems by improving reliability, robustness, and usability. This makes it a promising solution for practical healthcare applications, especially in environments where complete and accurate data may not always be available.

Acknowledgements

Authors are thankful to Sister Nivedita University for providing infrastructural support.

Declaration of Generative AI and AI-Assisted Technologies in the Manuscript Preparation Process

During the preparation of this work, the authors used AI-assisted tools ChatGPT-4 to enhance readability and improve language structure. These tools were not used to generate scientific content, analyses, or conclusions.

After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the final manuscript. No generative AI tools were used to create or modify the study figures or artwork

References

- Abbes, W., Frikha Elleuch, J. & Sellami, D. (2025). A new case based reasoning diagnosis approach within a possibilistic framework. *International Journal of Machine Learning and Cybernetics*, 16, 3297–3310.
- Aissaoui Ferhi, L., Ben Amar, M., Choubani, F., & Bouallegue, R. (2024). Enhancing diagnostic accuracy in symptom-based health checkers: a comprehensive machine learning approach with clinical vignettes and benchmarking. *Frontiers in Artificial Intelligence*, 7, 1397388.
- Ali, L., Rahman, A., Khan, A., Zhou, M., Javeed, A., & Khan, J. A. (2019). An Automated Diagnostic System for Heart Disease Prediction Based on χ^2 Statistical Model and Optimally Configured Deep Neural Network. *IEEE Access*, 7, 34938–34945.
- Aliyeva, K., & Mehdiyev, N. (2024). Uncertainty-aware multi-criteria decision analysis for evaluation of explainable artificial intelligence methods: A use case from the healthcare domain. *Information Sciences*, 657, 119987.
- Al-Nafjan, A., Aljuhani, A., Alshebel, A., Alharbi, A., & Alshehri, A. (2025). Artificial Intelligence in Predictive Healthcare: A Systematic Review. *Journal of Clinical Medicine*, 14(19), 6752.
- Al-qarni, S. S., & Algarni, A. (2025). Disease prediction from symptom descriptions using deep learning and NLP technique. *International Journal of Advanced Computer Science and Applications*, 16(5), 416-426.
- Azeem, M., Javaid, S., Khalil, R. A., Fahim, H., Althobaiti, T., Alsharif, N., & Saeed, N. (2023). Neural Networks for the Detection of COVID-19 and Other Diseases: Prospects and Challenges. *Bioengineering*, 10(7), 850.
- Chatterjee, V., Maitra, A., Ghosh, S., Banerjee, H., Puitandi, S., & Mukherjee, A. (2024). An efficient approach for breast cancer classification using machine learning. *Journal of Decision Analytics and Intelligent Computing*, 4(1), 32–46.
- Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J. J., Sandhu, S., Guppy, K. H., Lee, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American Journal of Cardiology*, 64(5), 304-310.
- Efunwoye, I. O., Gogate, M., Hussain, A., Luo, B., Ren, J., Jiang, F., Hussain, A., & Dashtipour, K. (2025). Human Disease Prediction Based on Symptoms Using Novel Machine Learning. In: Hussain, A., Jiang, B., Ren, J., Mahmud, M., Yang, E., Zheng, A., Li, C., Wang, S., Gao, Z., & Zhao, Z. (eds) *Advances in Brain Inspired Cognitive Systems (BICS 2024)*. Lecture Notes in Computer Science, vol 15498 (pp. 159-167). Singapore: Springer.
- Ferrante, M., Boccato, T., & Toschi, N. (2024). Enabling uncertainty estimation in neural networks through weight perturbation for improved Alzheimer's disease classification. *Frontiers in Neuroinformatics*, 18, 1346723.
- Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In: Balcan, M. F., & Weinberger, K. Q. (eds) *The 33rd International Conference on Machine Learning. Proceedings of Machine Learning Research*, vol 48 (pp. 1050-1059). New York, USA
- Geifman, Y., & El-Yaniv, R. (2019). SelectiveNet: A Deep Neural Network with an Integrated Reject Option. In: Chaudhuri, K., & Salakhutdinov, R. (eds) *The 36th International Conference on Machine Learning. Proceedings of Machine Learning Research*, vol 97 (pp. 2151-2159). Long Beach, CA, USA.
- Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On Calibration of Modern Neural Networks. In: Precup, D., & The, Y. W. (eds) *The 34th International Conference on Machine Learning. Proceedings of Machine Learning Research*, vol 70 (pp. 1321-1330). Sydney, Australia: International Convention Centre.
- Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In: von Luxburg, U., & Guyon, I. (eds) *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)* (pp. 6405-6416), Long Beach, CA, USA: Curran Associates Inc.

- Lisboa, P. J., & Taktak, A. F. G. (2006). The use of artificial neural networks in decision support in cancer: A systematic review. *Artificial Intelligence in Medicine*, 19(4), 408-415.
- Mohajer-Bastami, A., Moin, S., Ahmad, S., Ahmed, A. R., Pouwels, S., Hajibandeh, S., Yang, W., Parmar, C., Kermansaravi, M., Khalil, M., Khalid, A. W., Khamise, A., Rawaf, D., Hosseini, F., Agarwal, A., Lala, A., Ahmed, S., Patel, B., Fyntanidou, B., Egan, R., Mougiakakou, S. G., Jakob, D. A., Ribordy, V., Hautz, W. E., & Exadaktylos, A. K. (2025). Artificial intelligence in healthcare: applications, challenges, and future directions. A narrative review informed by international, multidisciplinary expertise. *Frontiers in Digital Health*, 7, 1644041.
- Noor, A. A., Manzoor, A., Mayhar Qureshi, M. D., Qureshi, M. A., & Rashwan, W. (2025). Unveiling Explainable AI in Healthcare: Current Trends, Challenges, and Future Directions. *WIREs Data Mining and Knowledge Discovery*, 15(2), e70018.
- Shortliffe, E. H. (1975). MYCIN: A Knowledge-Based Computer Program for the Diagnosis of Bacterial Infections. In *Proceedings of the Symposium on Computer Applications in Medical Care* (pp. 66-69). Las Vegas, Nevada: AMIA.
- Singh, V., Asari, V. K., & Rajasekaran, R. (2022). Deep Neural Network for Early Detection and Prediction of Chronic Kidney Disease. *Diagnostics*, 12(1), 116.
- Sogandi, F. (2024). Identifying disease symptoms and general rules using supervised and unsupervised machine learning. *Scientific Reports*, 14, 17956.
- Tasin, I., Nabil, T. U., Islam, S., & Khan, R. (2023). Diabetes prediction using machine learning and explainable AI techniques. *Healthcare Technology Letters*, 10(1–2), 1–10.
- Van Calster, B., Collins, G. S., Vickers, A. J., Wynants, L., Kerr, K. F., Barreñada, L., Varoquaux, G., Singh, K., Moons, K. G., Hernandez-Boussard, T., Timmerman, D., McLernon, D. J., van Smeden, M., Steyerberg, E. W., & Topic Group 6 of the STRATOS initiative. (2025). Evaluation of performance measures in predictive artificial intelligence models to support medical decisions: overview and guidance. *The Lancet Digital Health*, 7(12), 100916.
- Yang, Y., Zhang, G., Wu, Y., Zhao, Z., & Fu, Y. (2024). Adaptive Top-K Algorithm for Medical Conversational Diagnostic Model. *Entropy*, 26(9), 740.
- Yu, Z., Wang, K., Wan, Z., Xie, S., & Lv, Z. (2023). Popular deep learning algorithms for disease prediction: A review. *Cluster Computing*, 26(2), 1231-1251.