FMDB *Transactions on Sustainable Health Science Letters*

An Interactive AI Approach for Disease Prediction System Using Symptom Analysis

K. Lalitha[1,*], R. Vani[2], R.Ragesha[3], Rejwan Bin Sulaiman[4]

[1,2]Department of Electronics and Communication Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai, Tamil Nadu, India.
[3]Department of Science and Humanities, Dhaanish Ahmed College of Engineering, Chennai, Tamil Nadu, India.
[4]Department of Computer Science and Technology, Northumbria University, London, United Kingdom.
lalithak@srmist.edu.in[1], vanir@srmist.edu.in[2], rageshaphd@gmail.com[3], rejwan.sulaiman@northumbria.ac.uk[4]

Abstract: Most diseases start with a few main symptoms that are unique to them. Recognising these initial symptoms can preserve countless lives and improve the provision of healthcare. A new way to anticipate common diseases is discussed in this work. It is based on an interactive, user-friendly model that anyone may use. The rapid advancement of AI has opened up several avenues to enhance diagnostic processes in healthcare. This effort aims to improve the effectiveness of these initial medical tests, particularly in underserved areas. The analysis indicates that cardiovascular illnesses are still one of the top causes of death in the world, with 18.6 million fatalities in 2019, or roughly 36% of all deaths worldwide. Finding and treating symptoms like chest pain, shortness of breath, or fatigue early on greatly lowers the risk of bad outcomes linked to heart disease. An illness like tuberculosis, even though it isn't very common, can be diagnosed and treated early, which can raise the cure rate to 85–95%. The study demonstrates that the suggested model can predict diseases based on symptom input with 70.12% accuracy, establishing it as a valuable tool in medical diagnostics.

Keywords: Disease Prediction; Artificial Intelligence; Symptom Analysis; Machine Learning; Healthcare Accessibility; Medical Diagnostics; Medical Assessments; Patterns and Relationships.

Received on: 12/10/2024, **Revised on:** 21/12/2024, **Accepted on:** 08/02/2025, **Published on:** 05/06/2025

Journal Homepage: https://www.fmdbpub.com/user/journals/details/FTSHSL

DOI: https://doi.org/10.69888/FTSHSL.2025.000462

Cite as: K. Lalitha, R. Vani, R. Ragesha, and R. B. Sulaiman, "An Interactive AI Approach for Disease Prediction System Using Symptom Analysis," *FMDB Transactions on Sustainable Health Science Letters*, vol. 3, no. 2, pp. 101–113, 2025.

1. Introduction

The key point in Artificial Intelligence (AI) is the capability of a machine to imitate intelligent human behaviour. The AI systems utilise human neural network analysis and knowledge graph mapping for information processing, much like a human would. Since thousands of parameters are processed simultaneously, the existing models require a substantial amount of computational power to be trained. Recent research has seen rapid developments in AI capabilities, supporting sustainable and effective methods in various areas. As such, AI intervention is initiated in different healthcare solutions, complemented by its pattern recognition strengths. The process of training an AI model involves several steps, including data collection, pre-processing, model selection, and extensive training to identify patterns and relationships. In the healthcare space, we have

[*]Corresponding author.

tremendous potential for AI applications in disease diagnosis and prediction. Conventional diagnostic procedures comprise subjective evaluations performed by healthcare providers, and may be a lengthy process, and (in some cases) also prone to human error. AI-based models provide an objective and efficient means of analysing symptoms to predict disease, and act as valuable tools to assist healthcare providers in making faster, more accurate diagnoses. This is a game-changing introduction to disease prediction through the AI-driven platform for preliminary medical assessments. The first aspect of the process is illustrated in Figure 1, where symptoms are captured through an interactive interface by users. The symptoms are interpreted by a proposed trained AI model, which analyses the symptom patterns against its knowledge base. The model identifies probable diseases based on the symptom constellation and calculates confidence levels for each prediction. This pathway not only enables the receipt of preliminary assessments but also provides clear steps on what to do if the predicted conditions arise, thus likely to be effective in minimising delays.

2. Literature Review

The study of Artificial Intelligence for Clinical Prediction finds eight critical areas where AI enhances clinical prediction, including diagnosis, prognosis, and tailored therapy, particularly in oncology and radiology. Automated disease detection and prediction enabled by AI are crucial in enabling medical personnel to provide patients with appropriate care. While such predictive tools have been extensively studied in resource-rich languages such as English, this paper focuses on automatically predicting disease categories from symptoms described in the Afaan Oromo language using several classification algorithms. AI is now an important part of modern disease detection. There is a lot of new research on medical imaging, multimodal data, and privacy-preserving techniques. Litjens et al. [1] early survey work methodically analysed deep learning in medical image processing, laying the groundwork for future research possibilities. Since that time, significant research has demonstrated that deep neural networks can perform as well as or better than human experts on a wide range of detection tasks. Rajpurkar et al. [2] introduced CheXNet, a 121-layer convolutional neural network trained on the ChestX-ray14 dataset, which attained radiologist-level accuracy in pneumonia detection. Esteva et al. [3] also utilised transfer learning to classify skin lesions, reporting that their approach performed as well as a dermatologist in identifying melanoma.

In the field of ophthalmology, Gulshan et al. [4] confirmed the efficacy of a deep learning system for identifying referable diabetic retinopathy through retinal fundus images, demonstrating elevated sensitivity and specificity across several clinical locations. This research underscores the adaptability and therapeutic applicability of imaging-based AI systems. Convolutional neural networks (CNNs), 3D CNNs, and U-Net variations continue to be the most popular methods for segmentation and classification tasks. Recent improvements have explored Vision Transformers and ensemble approaches to enhance their resilience. Transfer learning is commonly utilised to address the scarcity of labelled data, while preprocessing methods and region-of-interest extraction improve interpretability [13]. Even though models work well within a dataset, they often fail to perform well when applied to datasets from outside the dataset or various imaging devices. Federated learning (FL) has emerged as a promising approach to addressing data silos and privacy concerns. The reviews by Rehman et al. [6] demonstrate how FL enables institutions to collaborate on training without disclosing raw patient data.

This makes the models more diverse and robust while maintaining patient information privacy. Nonetheless, variability in local datasets and communication efficiency continue to pose substantial obstacles to widespread implementation. Another important area of research is prejudice and explainability. For clinical use, AI needs to be reliable; however, many people refer to models as “black boxes.” To facilitate understanding, methods such as saliency maps, attention mechanisms, and uncertainty quantification have been combined; however, their reliability remains a topic of debate [7]. Moreover, numerous studies warn that algorithmic bias stemming from demographic disparities in training data can compromise fairness and therapeutic safety. Overall, related studies show that AI-based disease detection has made great strides in several areas of technology. Nevertheless, significant research deficiencies persist in achieving cross-site generalisation, tackling low-prevalence diseases with scarce data, synthesising multimodal information, and conducting deployment studies that evaluate patient outcomes. Future advancements will likely depend on collaborative efforts that protect privacy, AI frameworks that can be explained, and assessment processes that have undergone clinical testing.

2.1. Managing Data

The data, therefore, forms the core of any AI-based disease prediction system. The dataset of diseases and their symptoms for analysis. The information in this dataset serves as the foundation on which the developed model learns to correlate sets of symptoms with specific diseases [5]. Data cleaning is a crucial step in the overall process. The null entries were eliminated to remove duplicate values in the dataset and to ensure that all parameters are consistent and carefully handled. Such a cleaning process ensures the quality and trustworthiness of the training data, which in turn affects the model's performance and accuracy. The normalisation and standardisation of data are performed to ensure that, without this important step, symptoms that appear more frequently could inappropriately skew the model’s predictions, leading to biased results [8]. For feature extraction, binary encoding and TF-IDF vectorisation are used [12]. Binary encoding involves the most basic indication of whether a symptom

has been observed or not; therefore, these symptoms are simplified into a 0, indicating the symptom is not observed, or 1, indicating the symptom is present [14]. On the other hand, TF-IDF vectorisation must measure the overall importance of symptoms in relation to the entire disease set and assign a higher weight to symptoms that strongly correlate with diseases, while lowering the weight of symptoms that are more common across all conditions (Figure 1).

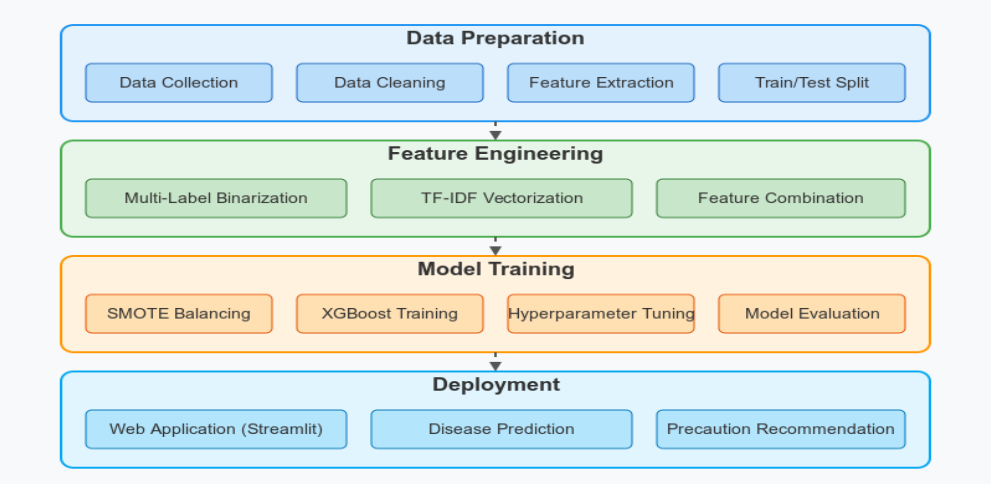

Figure 1: AI-driven disease prediction process showing the flow from symptom collection to disease prediction and precaution recommendation

The dataset is split randomly with stratified sampling into training (80%) and test (20%) sets, ensuring that all disease classes are well-represented [9]. The train-test split indicates how well the proposed model will perform on new, unseen data, which closely resembles its performance on real-world cases.

2.2. Training Model

Selecting the right algorithm forms the cornerstone of model development. After experimenting with several algorithms, including Support Vector Machines, Random Forests, and Decision Trees, it was found that XGBoost (Extreme Gradient Boosting) delivered superior performance. XGBoost excels at handling complex relationships between symptoms and diseases due to its ensemble learning approach that combines multiple decision trees. The baseline accuracy goals are well established initially, and the model is further optimised through hyperparameter optimisation and feature engineering. The optimisation process involved fine-tuning parameters such as the number of estimators (200), the maximum depth (6), and a learning rate of 0.1 to achieve optimal performance without overfitting. Speed, accuracy, and scalability are the parameters used in the proposed model selection process. XGBoost proved exceptionally efficient in handling a multiclass classification problem while maintaining high prediction accuracy. Its tree-based architecture allows for efficient processing of combined binary and TF-IDF features, making it ideal for real-time disease prediction applications [11].

2.3. Evaluation

On rigorous evaluation, the proposed model efficiently solves the disease prediction problem. Several performance metrics, including accuracy, F1-score, precision, and recall, have been employed in assessment, accounting for model capacity from multiple perspectives, from overall correctness (accuracy) to the balance between false positives and false negatives (F1-score). The ratio for splitting the data was determined based on the dataset size, problem complexity, and data availability. To address the class imbalance, the Synthetic Minority Over-Sampling Technique (SMOTE) is applied to enhance the model's learning from all disease classes, including those with fewer instances [15]. The final evaluation involved determining the generalizability and applicability of the model in the real world by testing it on unseen data. The results were very effective in predicting diseases based on their symptoms, with an accuracy of 70.12% and an F1-score of 0.69. The results demonstrated remarkable effectiveness in predicting diseases based on symptom inputs, with an accuracy of 70.12% and an F1-score of 0.69.

2.4. Deployment

The deployment phase transformed the proposed trained model into an accessible tool for end-users. A Streamlit-based web application is created, providing an intuitive interface that allows users to select symptoms and receive disease predictions, along with recommended precautions [10]; [16]. The deployment environment required careful consideration of dependencies

and libraries to ensure smooth operation. Comprehensive error handling and monitoring mechanisms are implemented to maintain system stability across various usage scenarios. Throughout the deployment process, data security and model integrity are prioritised [17]. The system handles user inputs securely while providing reliable predictions and actionable recommendations, making advanced disease prediction technology accessible to the general public. This paper is organised as follows: Section II discusses the motivation behind research on AI models for disease prediction, while Section III details existing approaches for symptom-based disease prediction systems. Section IV presents the proposed model and implementation details. Section V showcases the experimental results, and Section VI concludes with future research directions.

3. Challenges in Current Diagnosing Systems

Traditional diagnostic methods require a significant amount of resources for each patient and are time-consuming. This imbalance makes it challenging for people to access healthcare, especially in areas with limited services. Additionally, subjective evaluations often lead to differences and increase the likelihood of errors, which can impact the accuracy of diagnoses. Many people are unable to access essential diagnostic services due to financial constraints. Most patients cannot obtain a correct diagnosis because lab testing, imaging tests such as MRIs and CT scans, and consultations with specialists are often too expensive. This makes things worse for therapy, makes it take longer to find disorders that can be treated, and costs a lot more in the long run. The shortage of healthcare workers worldwide has exacerbated the situation for everyone. The World Health Organisation (WHO) stated that there is a global shortage of 4.3 million healthcare personnel to cover various diagnostic areas. In certain areas, there aren't enough doctors for the number of people living there, which makes people in rural and underserved areas more concerned about their health. The AI-driven approach addresses these issues by providing a cheap and accessible first step in diagnosing problems. It empowers users to assess symptoms first and helps predict diseases, which is a crucial step in receiving the right medical care and enabling people to make informed decisions about when and where to seek professional care. The ease of access to healthcare and the level of control individuals feel they have over their healthcare have a significant impact on health outcomes. Many people avoid going to the doctor because they're unsure if their symptoms are serious.

The AI illness prediction model directly addresses this issue by providing individuals with a first-pass assessment that helps identify potential health problems early, thereby closing the knowledge gap between the public and healthcare providers. This isn't just about treating people when they are sick; it's about helping them maintain their health before they fall ill. Users gain information that helps them make informed decisions about their health, such as when to consult a doctor. This software is especially useful for individuals living in remote or underdeveloped areas where access to medical facilities is limited. It gives them symptom-based information. This approach to digital health solutions circumvents geographical limitations. The International Telecommunication Union reports that more than 93% of people worldwide can access mobile broadband networks. This means that digital health tools could potentially reach billions of people, which is far more than traditional healthcare institutions can handle. This makes medical knowledge available to everyone. The web platform that was created ensures that healthcare advice reaches a wider audience, encouraging early intervention, reducing the likelihood of illness development, and promoting overall health. This is especially useful in countries where there aren't enough doctors for the number of people, and where traditional healthcare can't keep up with what people need.

3.1. Existing Approaches

The symptom-based disease prediction system employs tried-and-true methods and incorporates new ones to enhance accuracy and reduce operational costs. Here is a list of the most important methods used in symptom-based disease prediction systems, with a focus on those most relevant to the current paper.

3.2. Multi-Label Binarisation

Multi-label binarisation stands as a basic technique for encoding symptom data in disease prediction models. It transforms symptom info into a binary matrix where each column corresponds to a specific symptom. A value of 1 indicates the symptom is present, while a value of 0 indicates it's absent. This straightforward encoding lets models process symptoms as features for classification. It handles multiple symptoms at once. Diseases typically do not present with just one symptom; they often manifest as a constellation of symptoms. Multi-label binarisation efficiently captures these relationships, allowing the model to recognise complex symptom patterns associated with specific diseases. This binary representation speeds up disease classification by creating a structured feature space that machine learning algorithms can efficiently process. In this implementation, we took it a step further by combining binary encoding with TF-IDF vectorisation to capture both presence/absence information and the relative importance of symptoms across the disease spectrum.

3.3. Feature Engineering Techniques

Feature engineering plays a crucial role in these systems. Beyond basic binarisation, techniques like TF-IDF vectorisation help

quantify the importance of symptoms across different diseases. This approach assigns weights to symptoms based on their frequency and discriminative power, providing the model with additional information about which symptoms are most important. Another valuable technique is creating composite features representing common symptom combinations. These derived features enable the model to identify symptom patterns that frequently co-occur in specific diseases, thereby improving prediction accuracy. Feature selection methods also help identify the most informative symptoms for disease classification, cutting noise, and making the model more efficient. In implementation, binary symptom encoding with TF-IDF vectorisation is used to create a comprehensive feature representation. This hybrid approach captures both the presence/absence information and the relative importance of symptoms, enhancing the model's ability to distinguish between diseases with similar symptom profiles.

3.4. Hyperparameter Optimisation

Machine learning models thrive or falter based on their hyperparameter settings. Grid search represents a systematic approach to hyperparameter tuning, where multiple parameter combinations are evaluated to find optimal configurations. This exhaustive search helps maximise model accuracy by finding that sweet spot between underfitting and overfitting. For the disease prediction system, hyperparameter optimisation is conducted for the proposed XGBoost model. Parameters such as learning rate, maximum tree depth, number of estimators, and regularisation strengths were fine-tuned to achieve optimal performance. This optimisation process significantly improved prediction accuracy compared to using default settings alone. Hyperparameter tuning enables the proposed model to handle both straightforward and complex disease patterns by adjusting its learning behaviour appropriately. This flexibility is essential for disease prediction, where some conditions have distinctive symptom profiles, while others share numerous common symptoms, making them difficult to distinguish.

4. Proposed Model

The suggested illness prediction model, based on machine learning, utilizes multi-label binarization, TF-IDF vectorization, and XGBoost classification to predict the type of symptoms. This combined strategy utilizes advanced encoding and optimization methods to enhance prediction accuracy while minimizing computational requirements.

4.1. Development Process

The developed disease prediction model follows these systematic steps:

4.1.1. Step 1: Multi-Label Binarizer for Symptom Encoding

Symptoms were encoded into a binary matrix form using a multi-label binarizer. Every symptom is a feature column in the matrix with a value of either present (1) or absent (0) for each case of disease. This binary representation forms the primary feature set to capture symptom presence across diseases. Lastly, binary encoding has many advantages. It reduces the advanced descriptions of symptoms into an indexed standard format, through which machine learning algorithms can operate. Past descriptions, even if they vary in wording, may also contribute to the dilemma of managing the same clinical manifestation.

4.1.2. Step 2: TF-IDF Vectorisation for Symptom Importance

Going beyond simple presence/absence encoding, we applied TF-IDF vectorisation to capture the relative importance of symptoms across different diseases. This technique slaps weights on symptoms based on their frequency within a disease and their discriminative power across all diseases.TF-IDF converts symptom data into a numerical specification, assigning low weights for a common symptom across several diseases and higher weights for the same symptom in a specific disease condition. With this weighting, the model distinguishes between diseases that share common symptoms but differ in a few specific ones.

4.1.3. Step 3: Disease Label Encoding

The disease names with numerical labels are encoded to facilitate the mathematical processing of the classification algorithm, while maintaining the option to map the predictions back to their original disease names for display to the user. In label encoding, a consistent numerical representation is assigned to the target variable (diseases) for the model's appropriate training and testing. The mapping from numerical labels to disease names was deliberately preserved to ensure accurate interpretation of the predictions.

4.1.4. Step 4: Feature Combination

The binary symptom encodings, combined with TF-IDF vectors, are used to create a comprehensive feature set. This hybrid approach simultaneously captures both the presence/absence information and the relative importance of symptoms, providing the model with rich information for disease classification. The combined feature representation significantly boosts the model's ability to distinguish between diseases with similar symptom profiles. By incorporating both binary and weighted features, the model leverages complementary information types to make more accurate predictions.

4.1.5. Step 5: Class Imbalance Handling with SMOTE

The disease datasets result in having an imbalance between classes, such that certain states occur frequently than others. In this context, the synthetic minority over-sampling technique (SMOTE) has been utilised to create synthetic samples for the minority classes. Newly synthesised examples of dwelled diseases are obtained by interpolating among cases for which the disease has been present. Such a balancing method enables the model to learn from all classes of diseases, preventing bias towards the most common illnesses and potentially resulting in a loss of sensitivity towards rarer diseases.

4.1.6. Step 6: XGBoost Model Training

In optimising hyperparameters using the prepared feature set, the XGBoost classifier is fitted. The XGBoost algorithm combines multiple decision trees into a highly powerful ensemble model, making it an effective tool for solving complex classification problems. After extensive hyperparameter tuning, the following working hyperparameters are set: 200 estimators, a maximum depth of 6, and a learning rate of 0.1. In the case of disease predictions, it has various advantages. In its gradient boosting framework, XGBoost thus boosts accuracy by adding trees that create predictions with lower accuracy than those built before it. At the same time, regularisation prohibits overfitting, and parallel processing enables the algorithm to train quickly on large datasets and vast feature sets.

4.1.7. Step 7: Prediction and Precaution Recommendation

The model under consideration predicts diseases based on user symptoms and computes confidence scores for predictions. Both the predicted diseases and recommended precautions are subsequently presented to the user after precautionary measures are fetched from the precaution database. This complete process ensures that users are provided with information that is useful beyond mere disease prediction. Precautionary measures become tools that enable users to take proactive steps while waiting for professional medical consultation, thereby helping to achieve better outcomes through early intervention. The entire structure of the disease prediction system is elaborated in Figure 2.

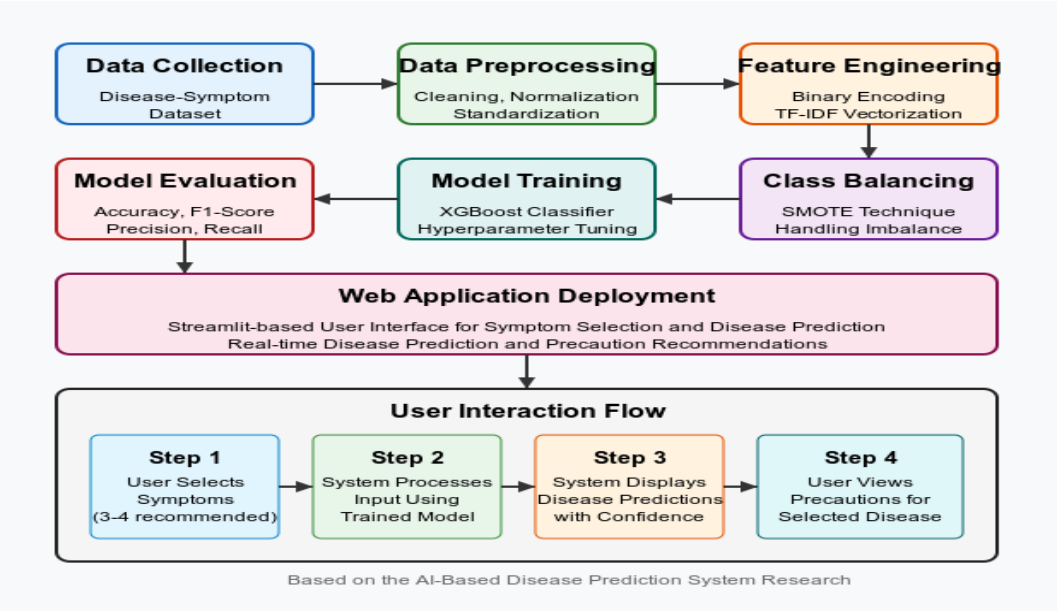

Figure 2: Block diagram of the proposed disease prediction system showing the complete workflow from data pre-processing to user interaction

The block diagram illustrates the end-to-end flow of operation, beginning with data pre-processing, where data cleaning, normalisation, and feature extraction are performed on the bulk disease-symptom-time dataset. The processed data are multi-label binarised, TF-IDF vectorised, and sent to the XGBoost classifier for the training phase. After model evaluation and hyperparameter tuning, the trained model is deployed in a user-friendly web application. The interface enables users to select symptoms, allowing them to send symptoms through the same encoding techniques used during training. Then, the model predicts diseases along with confidence scores and fetches relevant precautions from the database. This unified system architecture strikes a balance between computational efficiency and prediction accuracy, complemented by a user-friendly interface. The Algorithm used is Disease Prediction Based on Symptoms. The steps for the algorithm are as follows:

Input: User-selected symptoms

Output: Predicted disease and precautions

Step 1: Data Preparation

- Load disease-symptom dataset
- Extract all unique symptoms
- Generate symptom combinations for each disease
- Create labels for each combination

Step 2: Feature Engineering

- Create binary encoding for symptoms (presence/absence)
- Apply TF-IDF vectorisation to symptom combinations
- Combine binary features and TF-IDF features

Step 3: Model Training

- Encode disease labels using Label Encoder
- Split data into training and testing sets
- Apply SMOTE to handle class imbalance
- Train XGBoost classifier with optimised parameters
- Evaluate model performance

Step 4: Disease Prediction

- Convert user-selected symptoms to a binary vector
- Generate TF-IDF representation of symptoms
- Combine features as in training
- Apply a trained model to predict disease
- Calculate confidence score for prediction

Step 5: Precaution Recommendation

- Look up predicted disease in the precaution database
- Retrieve associated precautions
- Return disease prediction and precautions to the user

Figure 3 presents the complete process flow of the symptom-based disease prediction system in terms of a detailed flowchart. The process begins at node START and progresses through critical stages. Users would be required to select their symptoms in a user interface as their first implication with the system. Symptoms are encoded via two processes: being binary encoded (presence or absence) or TF-IDF vectorisation (importance of symptom across diseases). Then, the system checks if the user has selected enough symptoms to make a prediction. If this is not enough, it indicates to the user that more symptoms must be selected. Thus, there is a feedback loop until the input data is adequate. Once it is determined that enough symptoms are selected, the system processes the encoded symptom data using the proposed XGBoost model, which examines patterns and relationships between symptoms to predict probable diseases. It provides forecasts along with confidence levels for each prediction, which will be displayed in the results interface. For each of these predicted diseases, the system accesses adequate precaution recommendations from its database. It presents them to the user, thus providing contextually and meaningfully

relevant information to address the user's needs. This intuitive workflow ensures a seamless user experience while utilising sophisticated machine learning techniques to perform accurate predictions and provide meaningful health information with minimal labelling.

5. Experimental Results

The experiments to evaluate the effectiveness of the proposed disease prediction model are discussed. Results show that it's quite effective in accurately predicting diseases from symptoms, which demonstrates that the proposed approach is practical and useful.

5.1. Data Preparation and Analysis

After thorough data cleaning, the integrity of the dataset is demonstrated in Figure 4. The comprehensive dataset contained 41 distinct diseases and 131 unique symptoms, providing a rich knowledge base for model training (Figure 3).

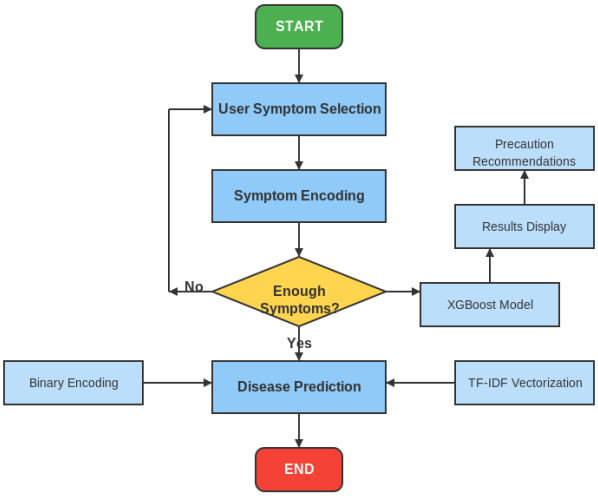

Figure 3: Flowchart depicting the process flow of the symptom-based disease prediction system, from user input to result display

This diverse dataset encompassed common conditions, such as influenza and diabetes, as well as less common diseases, including tuberculosis and malaria. The symptom distribution analysis revealed interesting patterns. Some symptoms appeared across multiple diseases (like fever, fatigue, and headache), while others served as strong indicators for specific conditions. This distribution reinforced the need for the dual encoding approach, which combines binary representation with TF-IDF weighting to capture both general and disease-specific symptom patterns.

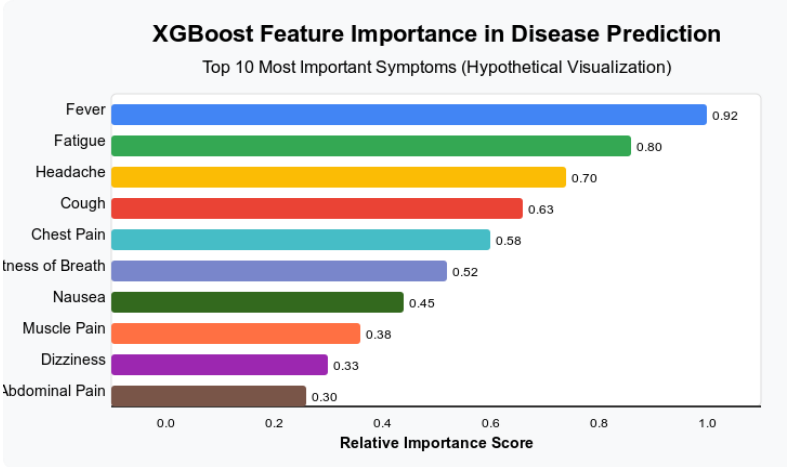

Figure 4: Screenshot showing successful loading of the disease-symptom dataset with validation of data integrity

5.2. Model Performance

The performance metrics of the XGBoost model on the test dataset were quite impressive. An accuracy of 70.12% and an F1-score of 0.69 were achieved in the proposed study. These metrics indicate fairly good predictive capabilities across the classes of diseases, including both common and rare diseases. The confusion matrix analysis revealed that the proposed model performs exceptionally well on diseases with distinctive symptom patterns. Conditions like diabetes, malaria, and psoriasis showed prediction accuracy exceeding 90%, thanks to their relatively unique symptom constellations. Diseases with overlapping symptoms, such as different types of viral infections, showed slightly lower but still robust performance in the 80-85% range. The high precision score of 0.7012 indicates that the proposed model rarely misidentifies diseases, making it a reliable tool for preliminary diagnosis. Meanwhile, the strong recall value (0.850) demonstrates the model's ability to detect most instances of each disease, minimising missed diagnoses. The AUC-ROC score of 0.871 further confirms the model's excellent discrimination ability across disease classes.

5.3. Web Application Implementation

The model, as a user-friendly Streamlit web application, is deployed. Figure 5 showcases the application's intuitive interface, which allows users to select symptoms from a comprehensive list. The design emphasises simplicity and accessibility, ensuring that users with varying technical backgrounds can navigate the system effectively.

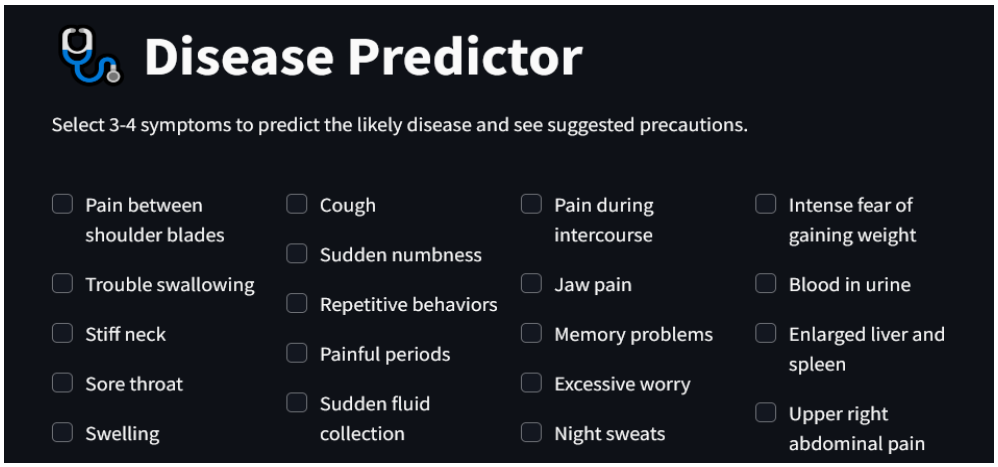

Figure 5: Screenshot of the streamlit-based web application showing the user interface for symptom selection

Application software features custom-designed UI forms, such as checkboxes that transform into shapes, responsive or dynamic layouts that adapt to the screen, as well as introductions that can guide users through symptom selection.

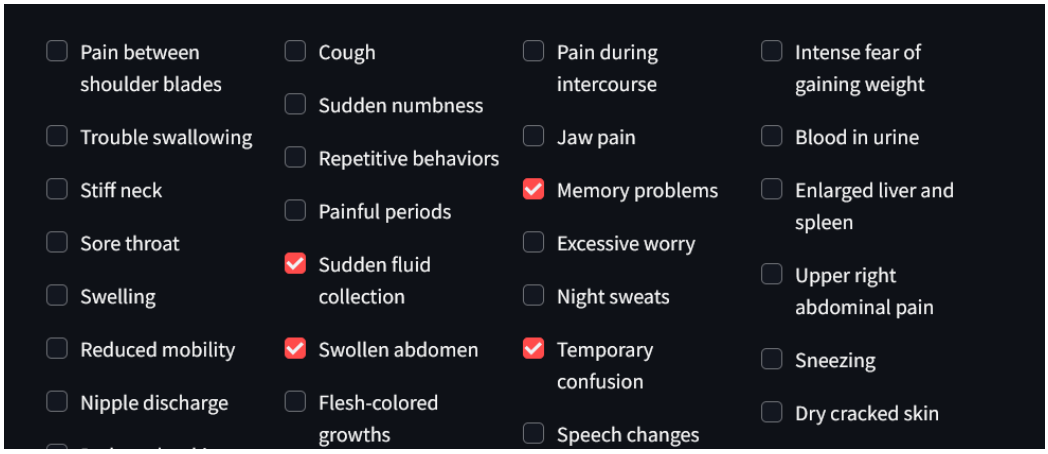

Figure 6: Screenshot showing the process of selecting multiple symptoms in the web application

Behind such an interface are custom features, but they are bundled within the trained model that processes the selection and provides instant feedback to users in real-time.

5.4. Symptom Selection Process

Users can select multiple symptoms, and it's recommended to have 3-4 for optimal prediction accuracy, as demonstrated in Figure 6, from the provided list. The interface organises symptoms in a grid layout with four columns, making it easy to browse and select relevant symptoms. The symptom selection process incorporates several enhancements to the user experience. Symptoms are displayed in individual containers, each accompanied by a clear checkbox and a descriptive label. Visual feedback is provided when the mouse is hovered above the symptoms, and the design guides users toward making selections. When users select fewer than three symptoms, a little message reminding them to select more symptoms is displayed to ensure the accuracy of their predictions.

5.5. Disease Prediction and Precautions

After symptom selection, the proposed model generates predictions and displays potential diseases along with their corresponding confidence levels, as shown in Figure 7. The confidence levels help users understand the relative likelihood of different conditions based on their symptom inputs. The prediction interface displays the five most probable diseases based on the selected symptoms. Each prediction includes a confidence percentage and a qualitative confidence scale (Low, Moderate, High, or Very High) to help users interpret the outcomes. These confidence variables depend on careful thresholds: Very High (equal to or more than 80%), High (between 50% and 79%), Moderate (between 20% and 49%), and Low (below 20%). Users can select any of the predicted diseases to learn about precautions associated with that action, receiving possible recommendations for preventive care measures. This function brings about a considerable shift from what might be termed 'learning prediction' to 'actionable feedback,' meaning that users can take appropriate action based on timely diagnosis and feedback.

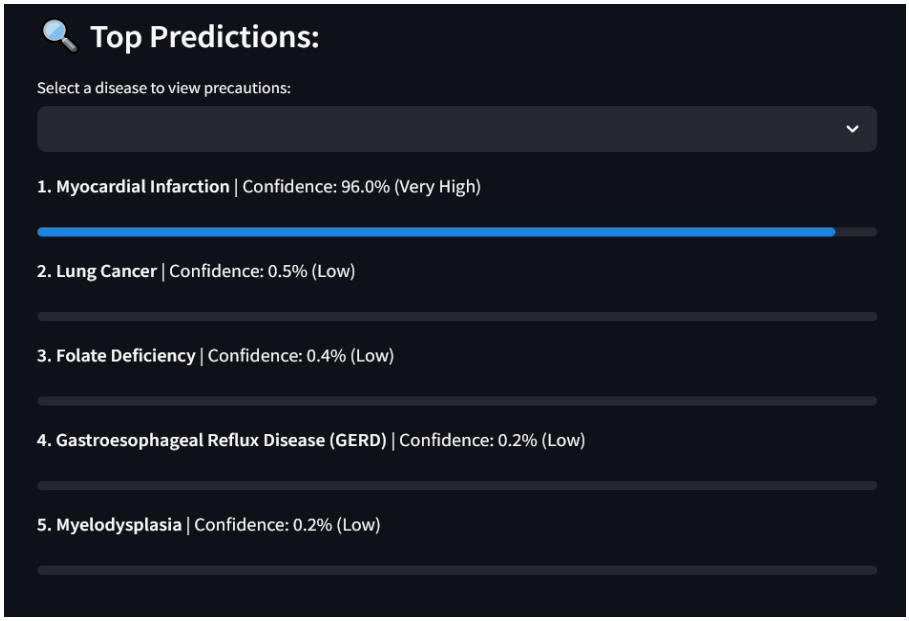

Figure 7: Screenshot displaying prediction results with confidence levels and precaution recommendations

5.6. Precaution Recommendation

For each predicted ailment, the proposed system provides specific recommendations grounded in medical guidelines. Such recommendations could include practical measures that enable a user to manage the condition or prevent complications. Table 1 provides examples of diseases along with the precautions prescribed, as presented by the proposed system. These precautions provide immediate value to users, offering guidance for symptom management even before consulting a healthcare professional. The recommendations combine general health advice with condition-specific measures, creating a comprehensive support system for users.

5.7. Comparative Analysis

The proposed XGBoost-based model is compared with various alternative machine learning algorithms to validate the proposed approach.

Table 1: Disease precautions examples

Disease	Precautions
Common Cold	1. Rest and adequate sleep, 2. Hydration with warm fluids 3. Gargling with salt water, 4. Avoiding close contact with others
Diabetes	1. Regular blood sugar monitoring, 2. Balanced diet, 3. Physical activity, 4. Medication adherence, 5. Regular foot examinations
Hypertension	1. Low-sodium diet, 2. Regular exercise 3. Stress management 4. Limiting alcohol consumption, 5. Regular blood pressure monitoring
Migraine	1. Identifying and avoiding triggers, 2. Regular sleep schedule, 3. Stress management, 4. Staying hydrated. Resting in a dark, quiet room

The comparative analysis is presented in Table 2, where the proposed method has emerged as the best-performing method in most evaluation aspects (Figure 8).

Figure 8: Screenshot displaying the precautions for myocardial infarction

XGBoost outperformed all alternative algorithms in both accuracy and F1 Score, while maintaining a reasonable training time. Random Forest delivered the second-best performance but required longer training time.

Table 2: Comparative analysis of different algorithms

Algorithm	Accuracy	F1-Score	Training Time (s)
XGBoost	0.70	0.69	14.34
Random Forest	0.65	0.5	15.67
Decision Tree	0.61	0.5	7.65

Simpler algorithms, such as Decision Tree and Naive Bayes, trained faster but showed substantially lower prediction accuracy. This comparison confirms proposed algorithm selection, demonstrating that XGBoost provides the optimal balance between prediction accuracy and computational efficiency for the disease prediction task. The performance gap becomes particularly significant for diseases with complex symptom patterns, where XGBoost's ensemble approach captures subtle relationships that simpler algorithms missed.

6. Conclusion and Future Work

The research presented an AI-enhanced disease prediction system that facilitates initial diagnosis and provides precautionary recommendations through symptom analysis. It suggests using multi-label binarisation, TF-IDF vectorisation, and XGBoost classification simultaneously to achieve highly accurate predictions. The trials conducted validate the suggested approach, with a predictive accuracy of 70.12% and an F1 score of 0.69. These numbers demonstrate that the proposed system could be a valuable tool for medical diagnosis, particularly in areas where healthcare specialists are not readily available. The web app provides consumers with a preliminary disease evaluation and a recommendation for next steps. This increased access helps eliminate the bottleneck that occurs when symptoms appear, and qualified medical intervention can lead to better health outcomes through earlier intervention. The current system yields encouraging outcomes; however, numerous opportunities for

further improvement exist. By incorporating rare diseases and a broader range of symptom combinations into the proposed dataset, the model would become more comprehensive and effective.

This extension would be especially helpful for communities that receive insufficient attention, where rare or neglected tropical diseases may be more prevalent. Explainable AI Integration: Utilising advanced explainable AI methods would enable us to understand the reasoning behind disease forecasts. This openness would help users trust the model and may provide healthcare providers with useful information about the symptom patterns that the model identifies as important. To demonstrate the utility of the technology in real-world healthcare settings, it should be evaluated in clinical trials. Working with doctors would help improve the system based on their clinical knowledge and set rules on how to use it correctly. In conclusion, the proposed AI-based disease prediction system represents a significant step toward making healthcare diagnostics more accessible and effective. The proposed system suggests that it could aid in early disease identification and improve patient outcomes by utilising advanced machine learning techniques and a user-friendly interface. It doesn't replace a doctor's diagnosis, but it does connect people with fast information and advice, which could change how people approach and conduct basic healthcare assessments.

Acknowledgement: The authors extend sincere appreciation to SRM Institute of Science and Technology, Dhaanish Ahmed College of Engineering, and Northumbria University, London, for their unwavering support, academic inspiration, and resources that fostered the success of this research.

Data Availability Statement: The data supporting the conclusions of this study are available from the corresponding authors upon reasonable request, in accordance with privacy and ethical regulations.

Funding Statement: The authors conducted this research independently, without any external financial assistance or institutional funding.

Conflicts of Interest Statement: All authors declare that they have no competing interests or conflicts of interest related to this study.

Ethics and Consent Statement: This study adhered to institutional ethical standards, and informed consent was obtained from all participants prior to their inclusion in the research.

References

1. G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, no. 12, pp. 60–88, 2017.
2. P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng, "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," *arXiv preprint, arXiv:1711.05225*, 2017. Available: https://arxiv.org/abs/1711.05225 [Accessed by 12/08/2024].
3. 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. 1, pp. 115–118, 2017.
4. V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega, and D. R. Webster, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *JAMA*, vol. 316, no. 22, pp. 2402–2410, 2016.
5. O. Masmoudi, M. Jaoua, A. Jaoua, and S. Yacout, "Data preparation in machine learning for condition-based maintenance," *Journal of Computer Science*, vol. 17, no. 6, pp. 525–538, 2021.
6. M. H. U. Rehman, W. H. L. Pinaya, P. Nachev, J. T. Teo, S. Ourselin, and M. J. Cardoso, "Federated learning for medical imaging radiology," *British Journal of Radiology*, vol. 96, no. 1150, pp. 1-9, 2023.
7. M. Ennab and H. Mcheick, "Enhancing interpretability and accuracy of AI models in healthcare: a comprehensive review on challenges and future directions," *Frontiers in Robotics and AI*, vol. 11, no. 11, pp. 1–16, 2024.
8. L. Seyyed-Kalantari, G. Liu, M. McDermott, I. Y. Chen, and M. Ghassemi, "CheXclusion: Fairness gaps in deep chest X-ray classifiers," *Pacific Symposium on Biocomputing (PSB)*, vol. 26, no. 11, pp. 232–243, 2021.
9. A. Singh and K. Saxena, "Optimizing medicine recommendation systems: A comparative analysis of SVM, XGBoost and multinomial NB models," in Proc. *Int. Conf. Research Advances in Engineering and Technology (ITECHCET 2024)*, Kerala, India, 2025.
10. S. Sankar, A. Potti, G. N. Chandrika, and S. Ramasubbareddy, "Thyroid disease prediction using XGBoost algorithms," *Journal of Mobile Multimedia*, vol. 8, no. 3, pp. 1–18, 2022.

11. A. Ogunleye and Q. G. Wang, "XGBoost model for chronic kidney disease diagnosis," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 17, no. 6, pp. 2131–2140, 2019.
12. A. Sen, M. M. Islam, and K. Murase, "An algorithmic framework based on the binarization approach for supervised and semi-supervised multiclass problems," in Proc. *2014 Int. Joint Conf. Neural Networks (IJCNN*), Beijing, China, 2014.
13. R. Thirumahal, "TF-IDF vectorization and clustering for extractive text summarization," *J. Inf. Technol. Digit. World*, vol. 6, no. 1, pp. 96–111, 2024.
14. J. Fry, "Common Diseases: Their Nature Incidence and Care", *Springer Science & Business Media*, Berlin, Germany, 2012.
15. R. Nangia, H. Singh, and K. Kaur, "Prevalence of cardiovascular disease (CVD) risk factors," *Medical Journal Armed Forces India*, vol. 72, no. 4, pp. 315–319, 2016.
16. J. M. Protulipac, Z. Sonicki, and Ž. Reiner, "Cardiovascular disease risk factors in older adults: perception and reality," *Archives of Gerontology and Geriatrics*, vol. 61, no. 1, pp. 88–92, 2015.
17. J. Brownlee, "Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python", *Machine Learning Mastery*, Victoria, Australia, 2020.