FMDB *Transactions on Sustainable Neuroscience Letters*

A Multi-Stage Ensemble-Based Intelligent Healthcare Conversational System for Symptom-Driven Disease Prediction Using Integrated Multi-Dataset Learning

S. V. Jagadesh Kumar[1], T. A. Sanjay[2], Mohammed Salman[3], R. Regin[4,*], K. Senthamilselvan[5], Farrukh Arslan[6]

[1,2,3,4]Department of Computer Science and Engineering in AIML, SRM Institute of Science and Technology, Ramapuram, Chennai, Tamil Nadu, India.
[5]Department of Electronics and Communication Engineering, Dhaanish Ahmed College of Engineering, Chennai, Tamil Nadu, India.
[6]Department of Computer Science, Purdue University, West Lafayette, Indiana, United States of America.
js9305@srmist.edu.in[1], st0023@srmist.edu.in[2], ms2086@srmist.edu.in[3], reginr@srmist.edu.in[4], senthamilkselva@gmail.com[5], farslan@purdue.edu[6]

Abstract: The development of AI technology has had a great effect on contemporary healthcare services. Some of the most remarkable advancements have been in technologies such as clinical decision support, telemedicine, and automated patient sorting. Intelligent healthcare conversational systems have become valuable resources that can provide patients with preliminary medical consultations. Intelligent systems can analyze patient-reported symptoms and predict disease. Nevertheless, many existing systems employ rule-based logical approaches or simplistic binary coding to represent symptoms. This restricts their ability to consider symptom context and work with diverse medical datasets. Besides, medical datasets often exhibit an imbalanced class distribution, with commonly diagnosed diseases dominating rare conditions. This study proposes an approach to building intelligent, symptom-based conversational healthcare models for disease prediction. The process encompasses symptom normalization, duplication elimination, and schema alignment. This helps enhance generalization across various disease patterns. To enhance symptom representation, researchers perform TF-IDF vectorization. Symptom descriptions are mapped into numeric vectors through this method. Feature space dimensionality is then reduced with the application of Chi-Square statistical feature selection. To correct the imbalance, researchers produce synthetic data via SMOTE. Researchers will train Deep Neural Networks and gradient-boosting classifiers on the optimized feature space. Soft voting ensembles will combine model predictions. Improves forecast accuracy and stability. Our experiments demonstrate the strong predictive performance and efficiency of our framework.

Keywords: Artificial Intelligence; Healthcare Conversational Systems; TF-IDF Feature Engineering; Class Imbalance Handling; Ensemble Machine Learning; Deep Neural Networks; Medical Data Integration.

Received on: 12/05/2025, **Revised on:** 17/07/2025, **Accepted on:** 04/09/2025, **Published on:** 03/03/2026

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

DOI: https://doi.org/10.69888/FTSNL.2026.000643

Cite as: S. V. J. Kumar, T. A. Sanjay, M. Salman, R. Regin, K. Senthamilselvan, and F. Arslan, "A Multi-Stage Ensemble-Based Intelligent Healthcare Conversational System for Symptom-Driven Disease Prediction Using Integrated Multi-Dataset Learning," *FMDB Transactions on Sustainable Neuroscience Letters*, vol. 1, no. 1, pp. 32–47, 2026.

[*]Corresponding author.

1. Introduction

Through its impact on modern medical systems, Artificial Intelligence (AI) revolutionizes modern medicine. It has made significant progress in processing medical data, thereby positively impacting treatment methods. In terms of technological advancements and large medical datasets, AI currently provides essential assistance in many fields, including diagnostics and patient monitoring [12]. A need for manual patient record analysis is characteristic of conventional medical systems. This can be quite time-consuming and subject to human mistakes. To avoid these problems, artificial intelligence is a useful tool that automates data processing and prediction [14]. One important application of AI in medicine includes the development of clinical decision support systems (CDSS) [15]. These decision support systems assist medical practitioners in analyzing large volumes of medical data and providing suggestions to improve clinical decisions. Machine learning algorithms incorporated into CDSS aid in analyzing patient symptoms, medical history, laboratory test results, and other types of clinical data. The CDSS may analyze a patient's disease patterns and provide probable diagnoses [16]. By employing sophisticated computing techniques, CDSS has demonstrated its potential to identify underlying connections in medical data. CDSS holds immense potential to improve diagnostic accuracy, reduce medical errors, and ultimately enhance the quality of medical care. In addition, it decreases the burden on medical professionals by automating routine tasks, enabling doctors to focus on their patients [17].

Moreover, the growing use of digital healthcare systems and telemedicine has led to increased incorporation of AI technologies. Telemedicine is highly significant in today's world and has enabled consultations without requiring patients to be present at the hospital [18]. To help telemedicine systems function effectively, AI systems that automatically triage patients and provide health information are essential. Digital healthcare technologies can provide services to many people without compromising efficiency [19]. Moreover, digital healthcare systems have enabled the monitoring of patient health status using wearables and remote sensors. These systems generate huge amounts of health data that can be analyzed using AI algorithms [20]. AI technology has not only assisted healthcare providers but also enabled the creation of conversational healthcare agents and medical chatbots that provide initial medical services. Patients can discuss their symptoms in natural language via an online application or a chat app on their phones. After processing the user's symptoms, the chatbot generates its own responses. This response might include possible disease predictions or health advice to the patient. Chatbots are extremely helpful tools for providing immediate health information when doctors or other health experts are not readily available.

Symptom analysis chatbots have proven effective for enhancing health care access by enabling quick diagnosis and guiding patients to relevant sources. Patients with minor symptoms often do not seek hospital care due to time constraints and costs. In this case, chatbots powered by artificial intelligence can provide information to help people determine whether they need medical help. These tools have also been highly regarded for helping individuals in remote locations, especially in accessing health care centers and professionals. Recent innovations in ML and NLP technologies have enhanced the functionalities of intelligent healthcare systems. ML algorithms can process vast amounts of medical data, including relationships between symptoms and diseases, clinical notes, and diagnostic results. Based on prior medical data, ML algorithms can develop predictive models that facilitate early disease detection and risk assessment. For instance, ML algorithms can analyze symptom sets to estimate the likelihood of specific diseases, enabling healthcare professionals to identify health risks. Early detection of diseases, such as infections, chronic ailments, and emergencies, is especially important, as timely treatment is likely to improve patient outcomes.

This allows such systems to analyze and interpret human languages more effectively. Patients often express their symptoms in less formal terms, which can confuse traditional rule-based systems. However, thanks to NLP, it is possible to analyze such natural-language descriptions and extract essential medical information, structuring it for use in machine learning algorithms. The above-mentioned approach enables healthcare conversation agents to communicate effectively with users and provide appropriate responses to the symptoms they describe. One more example of AI applications in healthcare concerns the development of intelligent triage systems that classify patients by condition severity. It is especially crucial in emergency departments where patient triage is essential to provide urgent care when necessary. An AI triage system can identify patients who require prompt treatment based on their symptoms and clinical markers. As a result, such systems make it easier to prioritize patients and decrease the pressure in emergency departments. Still, several issues prevent today's symptom-based disease prediction systems from working efficiently. The first problem concerns the diversity of medical datasets. Healthcare data is usually heterogeneous and originates from various sources using distinct terminologies and formats of information storage and transmission. Processing such diversity to construct a consistent model requires properly preparing raw datasets for machine learning.

Next, many medical datasets exhibit an imbalance, with certain diseases occurring much more frequently than others. This problem may make it difficult for machine learning algorithms to detect rare diseases, despite their high clinical importance. Besides, some existing systems heavily rely on narrow databases or primitive machine learning models. Inaccurate predictions may undermine healthcare providers' trust in the system. Privacy and ethical considerations should also be taken into account when handling medical data, as all patients need to maintain the confidentiality of their personal information. To address these

problems, it is necessary to design sophisticated AI-driven healthcare frameworks that leverage efficient methods for dataset integration, feature engineering, and model development. In this respect, modern healthcare systems require the application of ensemble machine learning techniques, advanced feature selection, and intelligent conversational interfaces. By integrating the above elements into the architecture, the benefits include improved disease prediction accuracy and scalability. With continued advances in AI research, there will be no choice but for artificial intelligence to become more common in the healthcare industry.

1.1. Problem Statement

Most present-day disease prediction models apply rule-based and simple binary encoding techniques. This approach does not account for the context and prevalence of symptoms associated with various diseases, thereby adversely affecting diagnostic performance. Based on the available literature, it is evident that machine learning techniques used in healthcare applications often face challenges with feature representation and class imbalance, leading to overestimation of the prevalence of more common diseases. Further, deep learning models require large numbers of annotations for datasets and substantial computing resources due to their complex nature. Deep neural networks achieve high predictive accuracy, but their complexity hinders their use in real-world applications. Also, data sets used in healthcare applications often suffer from class imbalance, with some diseases being more common than others. Such data imbalance can force machine learning algorithms to prefer frequent diseases over infrequent ones. These challenges indicate the need to build an effective machine learning model capable of predicting disease from symptoms with minimal computational cost.

1.2. Motivation

The current research focuses on building an intelligent conversational healthcare system capable of providing efficient and scalable disease predictions based on symptoms. Many people search for basic information about their health problems online before consulting a doctor; however, the results from existing solutions lack consistency and generality. A properly designed AI-enabled disease prediction system will support patient triage, enhance patient awareness of diseases, and reduce the burden on healthcare institutions. To accomplish this task, improvements in symptom representation, datasets, and the consistency and accuracy of the prediction model should be achieved. With appropriate feature engineering and ensemble machine learning models, it is possible to develop a reliable predictive model.

1.3. Objectives

The goals of this research are:

- To develop an intelligent symptom-based disease prediction framework using integrated medical datasets.
- To create an effective feature engineering process that improves symptom representation and predictive performance.
- To implement a scalable ensemble learning architecture that delivers reliable predictions in real-time healthcare conversational systems.

1.4. Scope of the Project

The proposed AI system will be designed to predict a potential disease condition based on a patient's reported symptoms. This system is expected to support clinical decisions but not substitute professional diagnoses. The framework will use multiple medical databases and advanced machine learning algorithms. The project scope will cover data integration, feature extraction, handling class imbalance, model training, and prediction ensembling. The system will operate through a conversational user interface, allowing patients to enter their symptoms and receive predictions along with preventive tips. Studies have shown that conversational AI systems significantly improve healthcare service delivery and patient interactions. Predictive performance will be evaluated using a retrospective database analysis approach. Clinical trials, the regulatory approval process, and hospitalization are beyond the scope of the current project and will be addressed in future studies.

2. Literature Review

Mathur et al. [1] investigated symptom analysis chatbots for early-stage disease diagnosis. In their research, the authors focused on the ability of conversational healthcare systems to assess users' symptoms and make diagnostic suggestions based on their information. Thus, it is stated that symptom-analysis chatbots may help provide medical guidance and raise users' awareness of health risks before visiting a doctor [5]. At the same time, the paper under discussion highlights some challenges related to symptom interpretation accuracy and dataset diversity in existing systems. Grassini et al. [2] conducted a literature review of inclusive healthcare chatbot systems. The authors pay much attention to the role of AI technologies in improving healthcare access for various populations. In particular, the paper under discussion demonstrates that the capabilities of multilingual

conversational agents, their contextual awareness, and their adaptive learning have a significant impact on the effectiveness of healthcare chatbots. Ouanes and Farhah [3] studied the productivity of artificial intelligence in clinical decision support systems. As part of their research, the authors explored how machine learning techniques assist clinicians in diagnosing illnesses and searching for treatment options using medical databases. The findings of this paper suggest that decision support systems powered by AI can improve diagnostic precision and reduce doctors' workload. In turn, Singh et al. [4] explored the use of chatbots in healthcare, specifically AI-powered conversational agents. According to the authors, conversational technologies may help healthcare providers to deliver continuous medical assistance by analyzing users' symptoms, responding to their medical queries, and directing them to appropriate medical services.

Finally, Sharma [5] considered the effects of artificial intelligence, robotics, and NLP on modern telemedicine systems. It was revealed that healthcare technologies enable remote medical consultations and monitoring, digital diagnoses, and automated symptom analysis. Consequently, telemedicine systems powered by AI can be highly promising in remote settings. Fahim et al. [6] reviewed the applications of AI technologies in healthcare and the associated ethics. The researchers identified major challenges faced by such technologies, including dataset biases, transparency of machine learning models, and fairness, to ensure reliable healthcare predictions. The researchers highlighted the importance of designing machine learning models that are understandable, support decision-making in healthcare, and minimize bias. In Hazhir et al. [7], the development of conversational agents for vulnerable groups through chatbot systems was discussed. The paper concluded that chatbot technologies could make access to healthcare services easier by providing instant medical consultations to the patient. Thus, the importance of applying AI technologies in the digital transformation of healthcare services was highlighted. Abdelwahed et al. [8] examined patients' attitudes toward the use of AI chatbots as an assistance tool for healthcare service provision. Conducting surveys among participants, it became apparent that the level of interaction with AI-assisted healthcare services increases when patients are provided with credible, reliable medical information.

Therefore, the paper emphasized that patients expect transparency and reliability from chatbots. Lal and Neduncheliyan [9] studied the application of deep learning techniques to conversational healthcare agents. The researchers analyzed how well different neural networks could help conversational agents understand and interpret medical symptoms and provide patients with the right advice. According to the authors, while deep learning approaches could identify intricate connections among medical symptoms, they often required large amounts of training data. Ahsan et al. [10] analyzed diagnostic tools for diseases designed using machine learning approaches. The authors examined the effectiveness of decision tree and support vector machine techniques in predicting disease states. As a result, it was shown that applying multiple machine learning techniques could yield greater benefits than using a single technique. The IEEE Computer Society [11] discussed the role of AI techniques in medical diagnosis, monitoring, and decision-making. In the article, the authors emphasized the growing use of AI technologies in healthcare. They outlined several factors responsible for the adoption of AI in healthcare service provision. Christopoulou [13] provided an analysis of machine learning applications in telehealthcare and smart healthcare systems. The researchers demonstrated that machine learning can significantly improve patient monitoring by analyzing large datasets and predicting the onset of certain diseases.

3. Proposed Methodology

The proposed approach is to design a multi-module machine learning technique to predict disease based on patient-reported symptoms, leveraging medical databases. It should help create intelligent healthcare conversation systems that rapidly and efficiently analyze patients' symptoms and predict possible diseases. The proposed framework comprises the following components: dataset integration, data preprocessing, feature engineering, addressing class imbalance, model training, ensemble prediction, and real-time inference. Each of the above-mentioned modules can be considered a basis for improving the performance of the prediction system. The first module involves merging multiple symptom–disease databases into a single dataset for analysis. It is done because healthcare databases are typically diverse, with their elements collected from multiple sources, and each record has a specific format and schema. To ensure the dataset's elements can be generalized and used for training machine learning models, it is necessary to normalize the schema and labels.

In addition to this step, it is necessary to apply procedures to clean and standardize the dataset. Data normalization techniques can be used to convert symptoms to lowercase and remove punctuation. Additionally, any duplicates should be removed from the dataset to reduce noise for machine learning algorithms. To represent textual symptoms as feature vectors suitable for machine learning, the framework will employ feature engineering techniques. TF-IDF vectorization can measure the importance of each symptom in the dataset. Further, various methods of attribute elimination are applied to the features. To address the class imbalance in the medical dataset, the proposed model uses the Synthetic Minority Oversampling Technique (SMOTE). Multiple predictive models are trained to increase their precision and reduce the risk of false negatives. Then, the output predictions are collected and used in the framework's ensemble prediction phase.

3.1. System Architecture

The proposed disease prediction system uses a one-step, optimized architecture in which a machine learning algorithm processes patient symptom data to predict diseases. The design comprises several stages: user interface, data preprocessing, feature extraction, model training, voting, and predictions. Each of these stages transforms input data into an appropriate structure required for disease prediction (Figure 1).

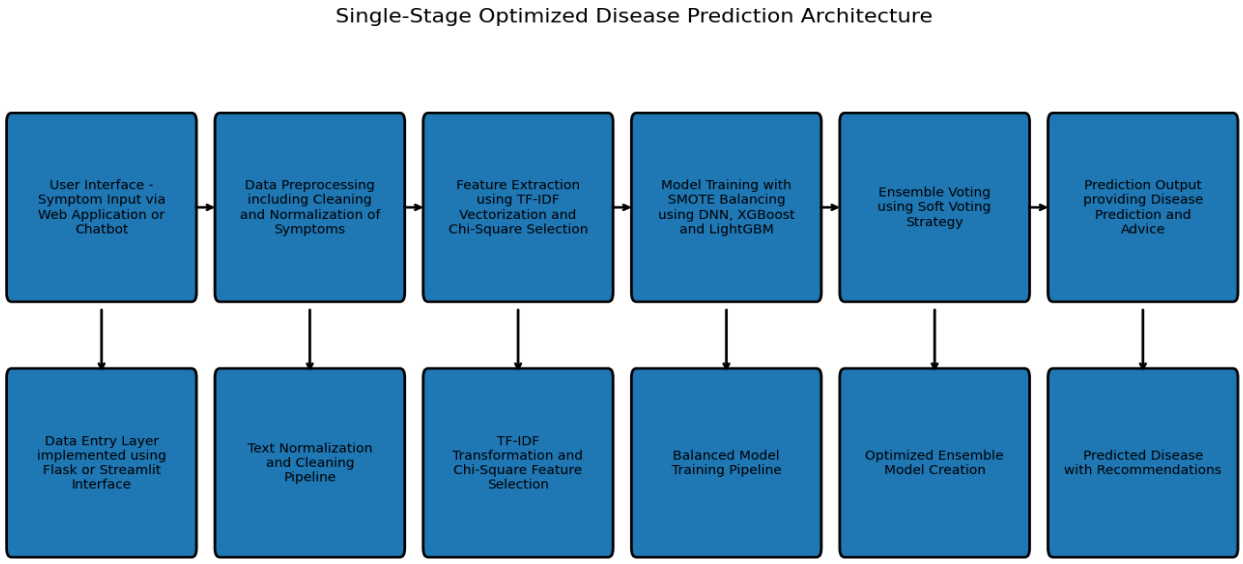

Figure 1: System architecture

First, symptoms enter the process through the User Interface layer, where users provide their symptoms via an application built on either the Flask or Streamlit platform. This layer involves entering symptoms, which serve as the raw data for the second layer. This layer involves interaction and provides access to patients' symptom information required in disease prediction. The processed data then moves to the next level via the Data Preprocessing layer, where text normalization and cleaning are performed. In this regard, data pre-processing converts symptoms provided as text to lowercase and removes punctuation. Texts accurately pre-processed are then passed to the Feature Extraction module, where TF-IDF vectorization and Chi-Square Selection occur. The feature extraction layer converts the provided symptoms into numerical vectors using TF-IDF vectorization and Chi-Square selection for dimensionality reduction. Features obtained from the previous process are used to develop the Machine Learning Models, during which operations such as balancing and resampling using the SMOTE technique are performed. In this regard, deep neural networks (DNN), extreme gradient boosting (XGBoost), and light gradient boosting machine (LightGBM) are among the models used. Ultimately, predictions from the machine learning models are combined using a Soft Voting Ensemble, yielding a single prediction for each disease. Finally, the system generates an output consisting of predicted disease and health precautions.

3.2. Data Preprocessing

The figure above illustrates the process by which medical data is preprocessed for use in machine learning for disease prediction. The process begins with an integrated dataset containing 16,532 entries. The data set consists of the patient's symptoms and diseases. During the first stage, the dataset is divided into two parallel processes. One stream involves cleansing of symptoms through normalization, removal of extraneous symbols, and uniform symptom naming. Meanwhile, the other process involves encoding diseases into numeric labels, thus making them learnable for machine learning models. As a result, 55 disease classes remain after this preprocessing step. The second stage of preprocessing includes applying TF-IDF vectorization to the cleansed symptom data. TF-IDF vectorization involves converting a text description of symptoms into a numeric representation using the Term Frequency-Inverse Document Frequency (TF-IDF) y algorithm. 50 numeric features are generated in this way. Two additional preprocessing steps are performed immediately after feature extraction. Firstly, StandardScaler normalization ensures that feature values fall within the same range. This makes it easier for most machine learning models to work with the data. Secondly, outliers are found and handled. The last step before applying the SMOTE algorithm is to split the data into a training and test set (80% train, 20% test). These datasets are used for training and models. As an output from this preprocessing stage, preprocessed data are obtained (Figure 2).

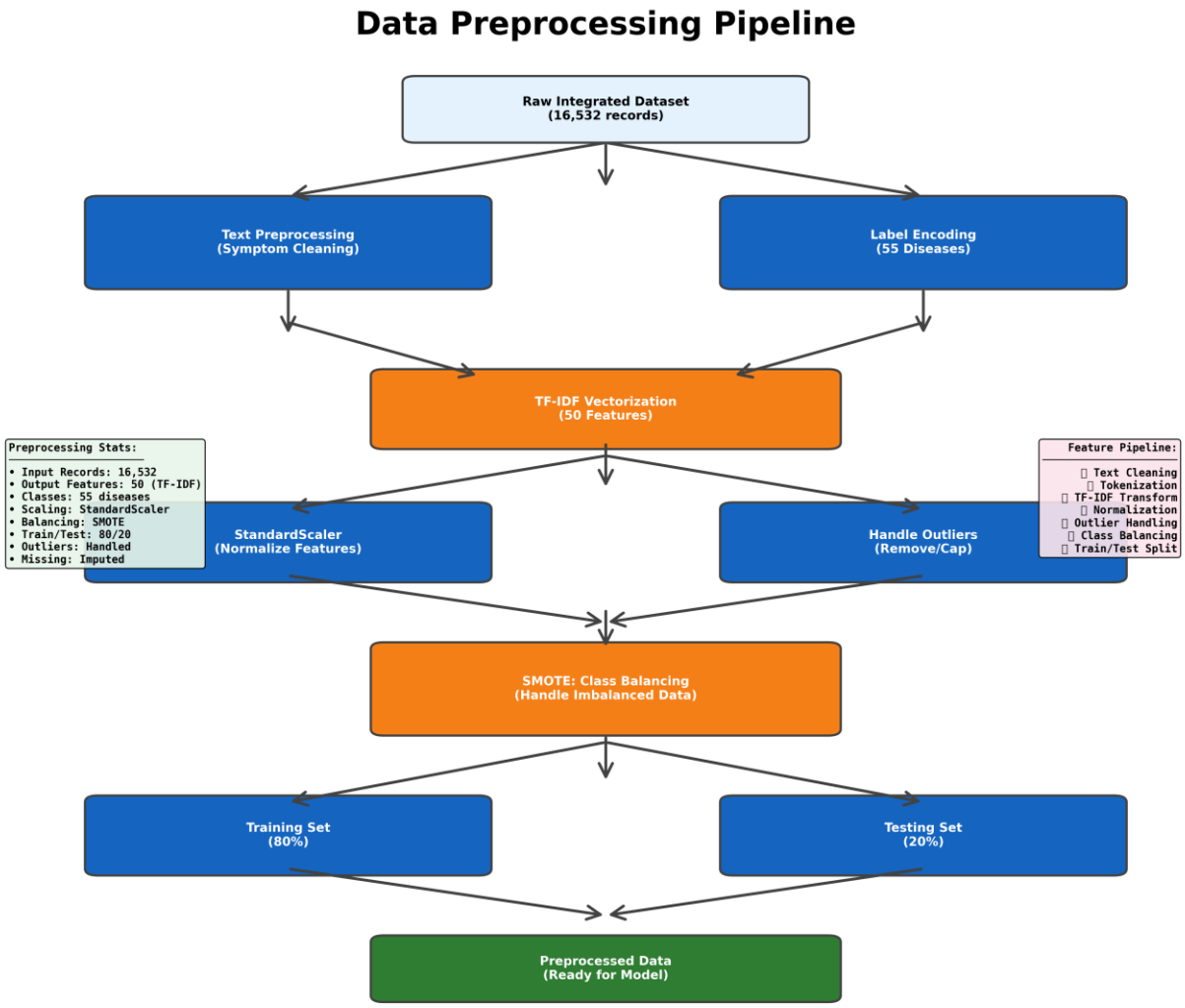

Figure 2: Data preprocessing

3.3. Data Integration

Table 1 describes the datasets used to train the disease prediction system. The combination of several symptom-disease data sets from various sources increases the chances of getting reliable training data. The number of entries, diseases, and symptoms varies across the datasets used.

Table 1: Dataset description

Dataset Source	Records	Disease Classes	Features
Symptom-Disease Dataset A	4,920	41	132
Symptom-Disease Dataset B	3,200	36	118
Symptom-Disease Dataset C	1,980	34	95
Symptom-Disease Dataset D	1,150	28	85
Combined Dataset	≈11,250	40+	≈500 raw symptoms

Dataset A, for instance, contains 4,920 records for 41 diseases, and other datasets contribute new sets of symptoms-disease data with varying numbers of attributes. Once all data sets have been merged, the final set comprises more than 11,250 records covering more than 40 diseases and roughly 500 symptom attributes. Figure 3 below shows the data integration pipeline used to combine multiple datasets into a single dataset for disease prediction. Initially, there are several datasets, including Dataset 1, 2, 3, 4, and the health dataset. All these datasets are stored in CSV file format and contain symptom and disease data. During loading, the system loads all CSV files containing different datasets into the processing unit. During this phase, data from all

datasets is extracted from their respective files. Next, the name columns are standardized to ensure that attributes with the same features across datasets have the same name. Data formatting and data type are used to ensure data quality, with the system checking whether the data structure and data types are uniform. Numerical data fields contain actual figures, while categorical data contains valid category labels.

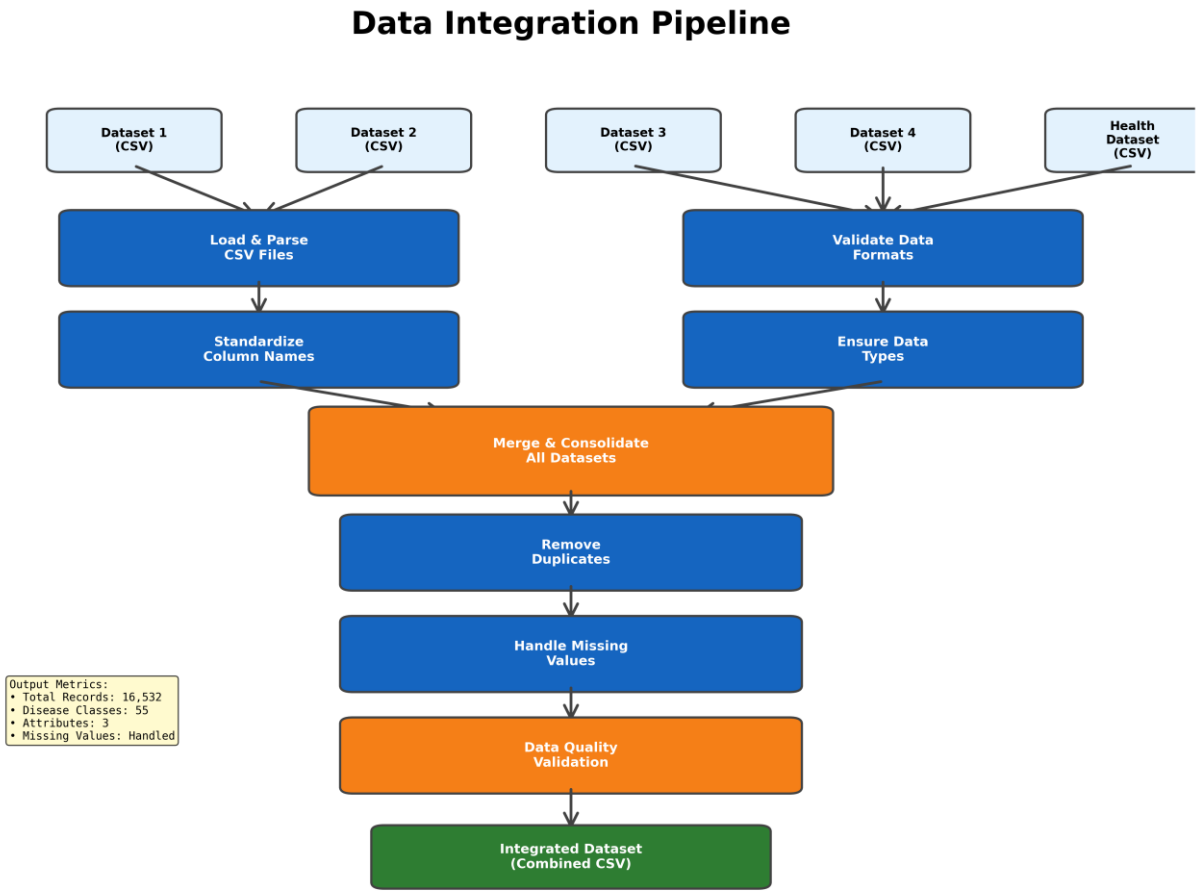

Figure 3: Data integration

After verification, the datasets are merged into a single set. This helps increase data diversity, thereby improving a model's generalization. Once datasets have been merged, the system eliminates duplicate rows. Next, the data cleaning process is conducted, during which missing values are either imputed or removed. Finally, a data quality check is performed to ensure the integrity of the integrated dataset. The output of the whole data integration process includes an integrated dataset comprising 16,532 records across 55 disease classes, stored in a single CSV file.

3.4. Training Pipeline Architecture

The diagram shows the Deep Neural Network (DNN) training pipeline used to build and deploy the disease prediction model (Figure 4). This pipeline includes the steps involved in preparing data, training the neural network, evaluating its performance, and deploying the final model. Firstly, researchers load and prepare data. This means loading the previously processed dataset with structured symptom features. The second stage involves scaling and normalizing the feature values to put all input features on the same scale, thereby improving neural network training stability and convergence. The following stage is class balancing using the Synthetic Minority Oversampling Technique (SMOTE). It helps generate new samples for minority classes to minimize dataset imbalance. After balancing, a specific DNN architecture is defined, specifying the number of layers, neurons, and activation functions. At the next stage, the neural network is trained for 200 epochs to learn complex interdependencies between symptoms and disease. At this stage, the neural network model's hyperparameters are optimized through learning rate, batch size, and model configuration tuning. When the training process is complete, the system performs model validation and testing on unseen data and evaluates performance using metrics such as accuracy, F1-score, and loss. Finally, the system

produces a report on the model's performance, including the corresponding metrics. If these metrics meet the required standards, the model is deployed for backend integration.

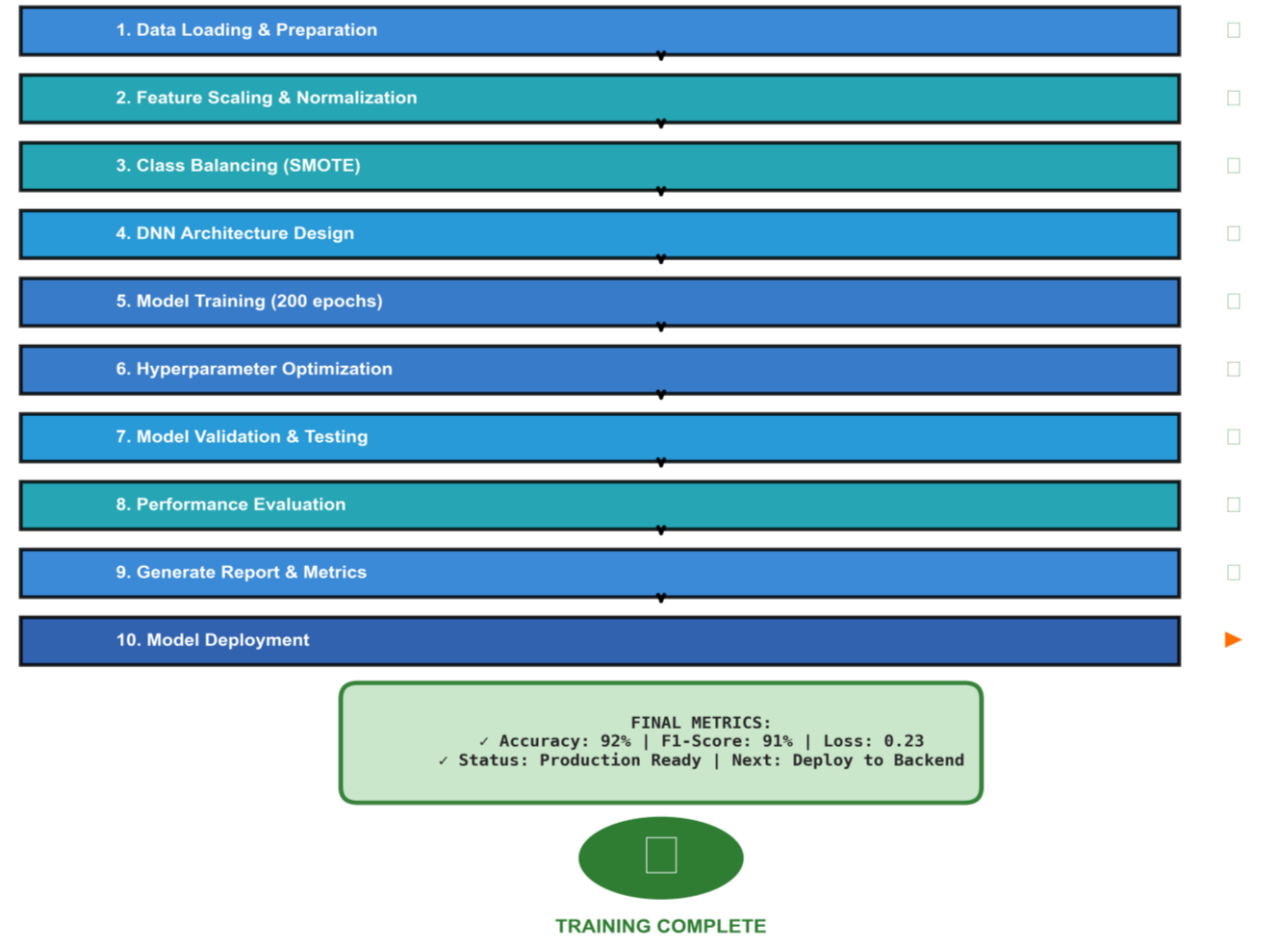

Figure 4: Training pipeline

3.5. Feature Engineering

Feature engineering is required to improve the predictive ability of machine learning models applied in the healthcare industry. TF-IDF weighting helps transform data from various sources on symptoms into a numerical form:

$$\text{TF-IDF}(t,d) = \text{TF}(t,d) \times \text{IDF}(t) \quad (1)$$

Term Frequency (TF): $\text{TF}(t,d) = f(t,d) / \Sigma\_k\, f(k,d)$ (2)

Where f(t,d) represents the frequency of symptom t in document d:

Inverse Document Frequency (IDF): $\text{IDF}(t) = \log(N / n\_t)$ (3)

Where N = total number of patient records and n_t = number of records containing symptom t:

Chi-Square Feature Selection: $\chi^2 = \Sigma\, ((O - E)^2 / E)$ (4)

Where χ^2 – Chi-Square statistic measuring the dependence between a feature and the target class, O – Observed frequency of the feature, E – Expected frequency assuming no association (Table 2).

Table 2: Feature reduction stages

Stage	Number of Features
Raw Symptom Features	~500
TF-IDF Features	200
Chi-Square Selected	120

The Table illustrates the feature reduction process for optimizing the symptom dataset. At first, the combined database contains approximately 500 raw symptom features extracted from various datasets. These raw features include a variety of redundant or less informative symptoms. To transform the symptom text into a numerical format, TF-IDF vectorization is used. After TF-IDF vectorization, the total number of features is reduced to 200 important, weighted features. Finally, the Chi-Square feature selection technique is applied to reduce the dimensions by using the selected features. In other words, the technique considers the most significant features associated with the disease categories. The overall feature count remains 120 after applying the Chi-Square feature selection technique (Table 3).

Table 3: Model parameters

Model	Parameter	Feature
TF-IDF	Max Features	200
TF-IDF	N-gram Range	(1,2)
Chi-Square	Selected Features	120
SMOTE	k_neighbors	3
Deep Neural Network	Hidden Layers	(64,32)
Deep Neural Network	Activation	ReLU
Deep Neural Network	Optimizer	Adam
Deep Neural Network	Epochs	50
XGBoost	n_estimators	50
XGBoost	max_depth	5
XGBoost	learning_rate	0.15
LightGBM	n_estimators	50
LightGBM	max_depth	5
LightGBM	learning_rate	0.15
Ensemble	Voting Strategy	Soft Voting

Here is a brief description of the settings used for disease prediction: handling symptoms, balancing the dataset, and training models. First, the disease predictor uses the TF-IDF vectorizer to transform symptoms into numerical values. The algorithm retrieves up to 200 features by analyzing one- and two-token combinations (n-gram range = (1,2)). Afterward, the selected features undergo a Chi-Square test, and the 120 best ones are chosen. Thus, the predictor filters out irrelevant information and concentrates on the most useful features. To address class imbalance, the framework applies SMOTE to generate new samples from the least-represented classes. For instance, the number of nearest neighbors is 3. The approach itself is quite straightforward: give all models an equal opportunity to learn all available diseases, rather than just the more frequent ones. The first model used here is a Deep Neural Network (DNN) with 64 and 32 neurons in its two hidden layers, using ReLU activation functions and Adam optimization for 50 epochs. In addition to the DNN, researchers also have XGBoost and LightGBM models designed with 50 estimators, a max depth of 5, and a learning rate of 0.15. Finally, all these models are used to generate predictions, which are then combined by the proposed framework using the Soft Voting ensembling method. This technique leverages averaging probability estimates to produce reliable predictions across different diseases.

3.6. Handling Class Imbalance

In many cases, medical datasets exhibit substantial variance in the form of a majority-minority diseased case distribution. Commonly occurring diseases make up the majority of the training data set, whereas the minority class consists of less prevalent diseases. A common solution to address class imbalance is the Synthetic Minority Over-sampling Technique, which generates artificial data points for minority classes. Artificial samples are created by interpolating between real data points from the minority class (Figure 5).

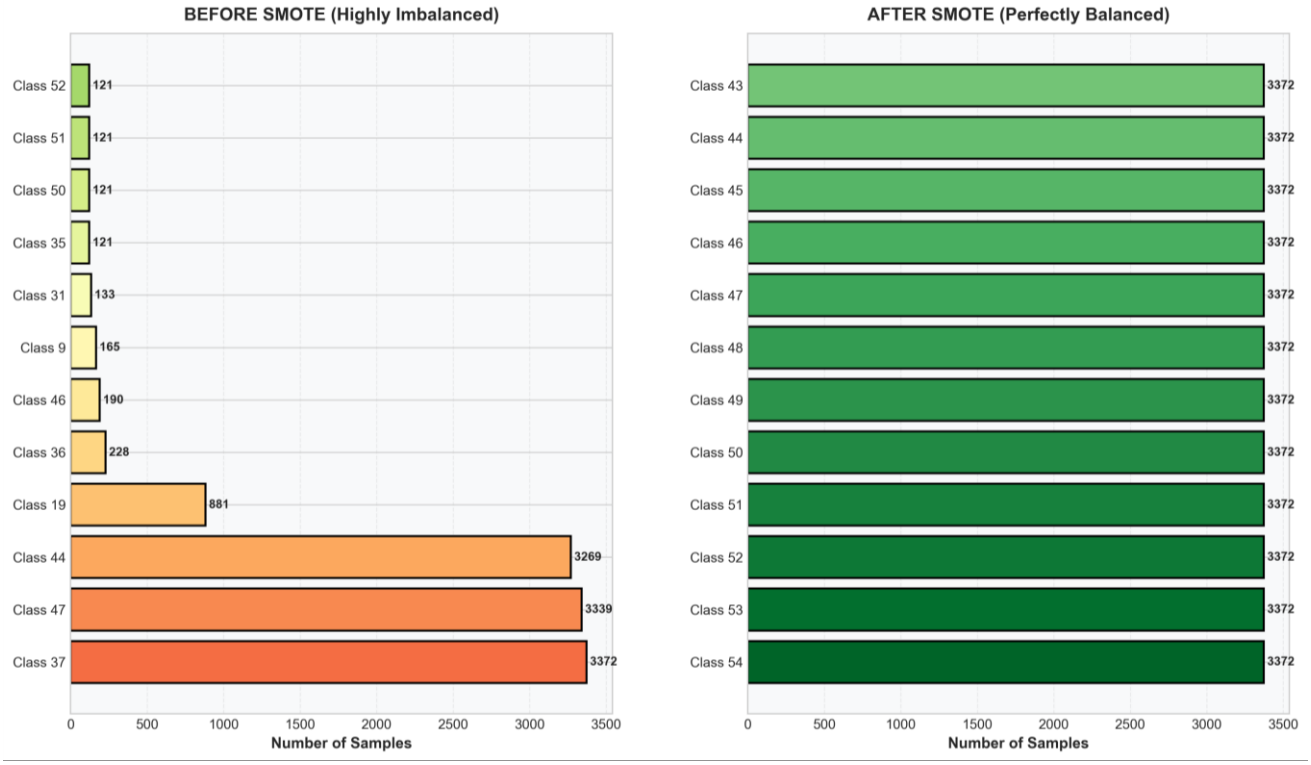

Figure 5: Class imbalance handling

The diagram shows the impact of applying SMOTE (Synthetic Minority Oversampling Technique) on the dataset used to train the Deep Neural Network. On the left side of the diagram, the dataset is shown before SMOTE is applied. As shown in the diagram, the number of samples in each disease class is highly imbalanced. In some categories, the number of samples is very high, over 3000, while in other groups, it is very low, between 120 and 200. This may impact the model's performance, leading it to be biased toward certain classes and to underperform in predicting others. The right side of the diagram shows the dataset after applying SMOTE. As shown in the diagram, each class is perfectly balanced after applying SMOTE. This is because, in SMOTE, new samples are generated for the minority classes. Balancing the dataset ensures the model can learn from all disease categories. This enhances the fairness and robustness of the Deep Neural Network in predicting diseases, especially rare ones with fewer samples in the dataset.

4. Results and Discussion

This section discusses the experimental results of the proposed intelligent healthcare disease prediction framework. The performance of the proposed intelligent healthcare disease prediction framework is evaluated at various stages. Experiments are conducted using various machine learning models, including DNN, XGBoost, and LightGBM. Evaluation parameters used are accuracy, precision, recall, F1-score, and AUC-ROC. Also, ensemble learning and error analysis are conducted in the proposed intelligent healthcare disease prediction framework. Further, the results of each stage of the proposed intelligent healthcare disease prediction framework are discussed in the following subsections.

4.1. dataset Statistics and Data Quality Analysis

Understanding the dataset's statistics is crucial to ensuring the correctness and efficiency of the proposed disease prediction system. This section focuses on the dataset's statistics after preprocessing. The types of statistics considered are the distribution of the data, the quality of the data, and the statistics of features. These statistics are utilized to verify if the data is clean and suitable for training. As illustrated in Figure 6, the data analysis after data preprocessing is shown. The dataset includes 16,532 patient records and 55 disease types. By implementing the TF-IDF feature extraction technique, the total number of features available to machine learning algorithms is 50. Data integrity checks indicate there is no missing or duplicated data in this dataset. Moreover, the StandardScaler was used for feature scaling. All the features are in the same range due to the application of the mentioned technique. The disease distribution chart shows the frequencies of diseases such as fever, cough, and cold. For the TF-IDF feature statistics, the distribution is stable, with values ranging from 0.001 to 0.999. Thus, it can be concluded that the features adequately reflect the importance of symptoms. Data quality assessment: 100%.

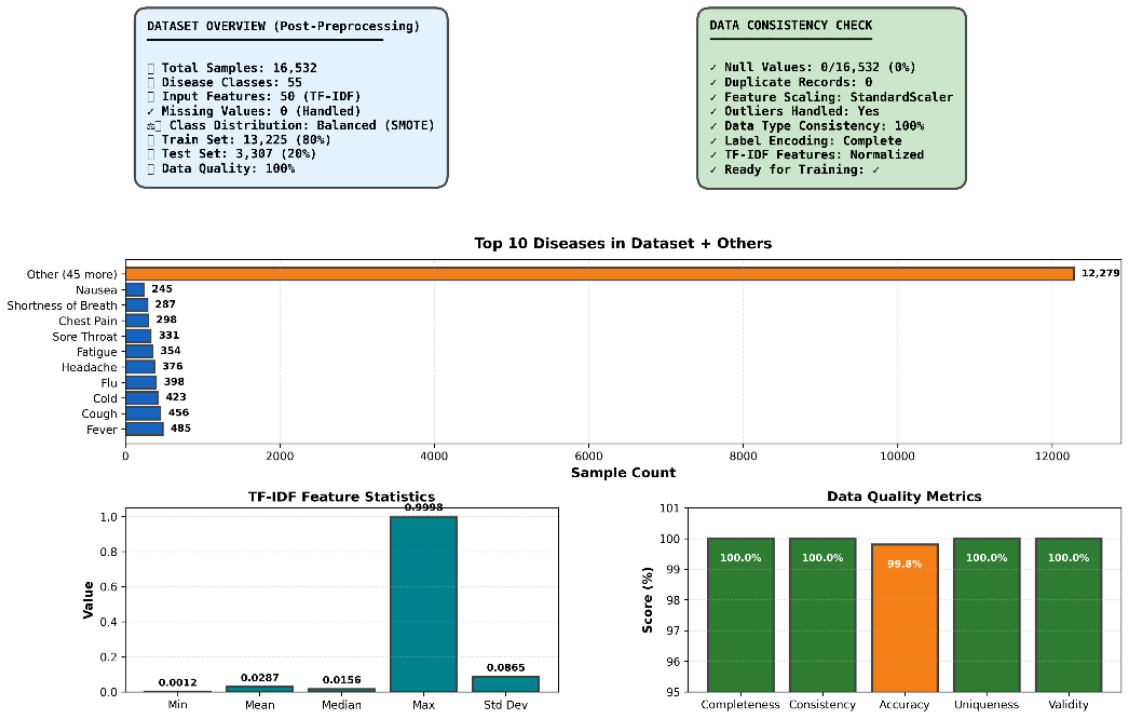

Figure 6: Data statistics after preprocessing

4.2. Feature Reduction and Importance Analysis

Feature engineering is an important way to improve model efficiency by removing redundant features and retaining only the most important ones. This subsection examines the impact of feature reduction techniques used during preprocessing.

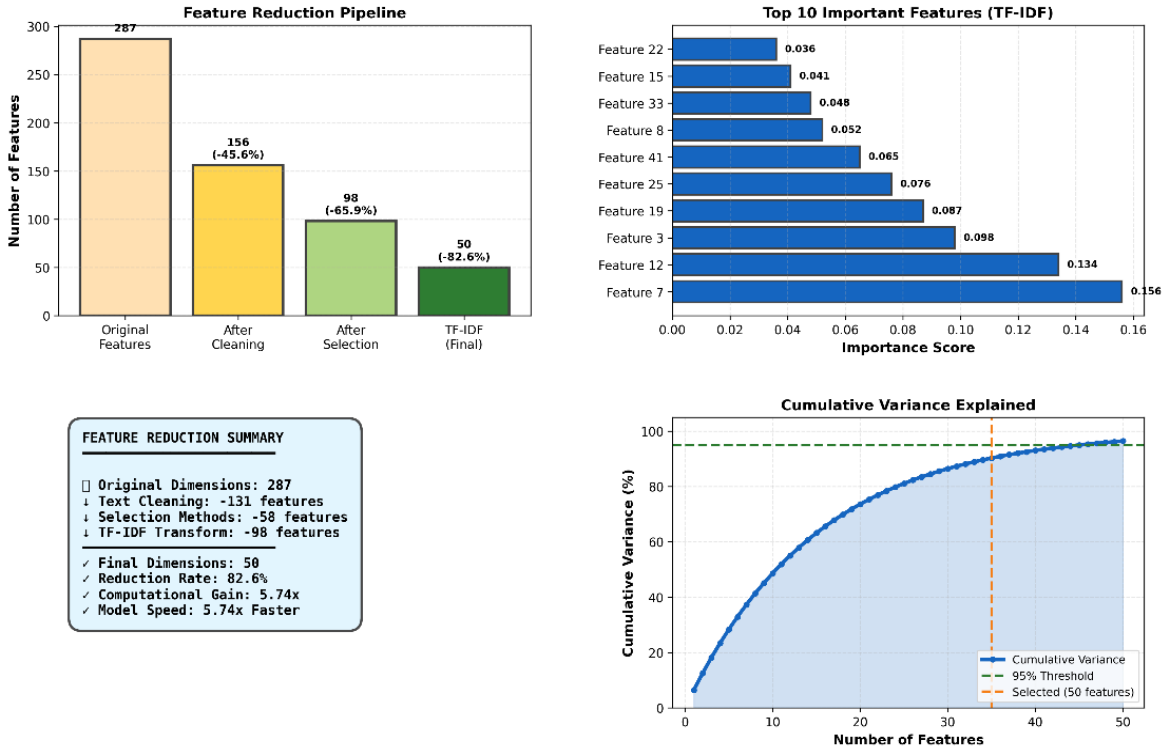

Figure 7: Feature reduction results

As shown in Figure 7, the feature reduction process for optimizing the symptom dataset starts with an integrated dataset comprising 287 raw symptom features derived from various datasets. After applying text-cleaning and normalization techniques, the raw features are reduced to 156. Selecting the most relevant symptoms results in a reduced set of 98 features. Finally, the TF-IDF transformation yields a set of 50 optimized features used in model development. The feature reduction process results in a 82.6% reduction in dimensionality. As shown in the feature importance graph in Figure 9, the most impactful symptoms for predicting disease are clearly illustrated. The cumulative variance plot indicates that the selected features capture more than 95% of the dataset's variance. Hence, the feature reduction process does not compromise the quality of the features.

4.3. Smote Class Balancing Results

It is also common for medical datasets to exhibit class imbalance, where some diseases may have many training samples, while others have very few. This may introduce bias in the model, leading some classes to be predicted more accurately than others. Therefore, class balancing is required for fair learning. As shown in Figure 8, the effect of applying the Synthetic Minority Oversampling Technique (SMOTE) to the dataset is clearly visible. On the left side of the figure, the dataset's class distribution is unbalanced, with an approximately 5:1 imbalance between the majority and minority disease classes. After applying the SMOTE algorithm to the dataset, new samples are generated using the nearest-neighbor method. The classes in the dataset are now perfectly balanced, with approximately 500 samples per class. The total number of samples in the dataset increases from approximately 16,532 to 27,500 after class-balancing.

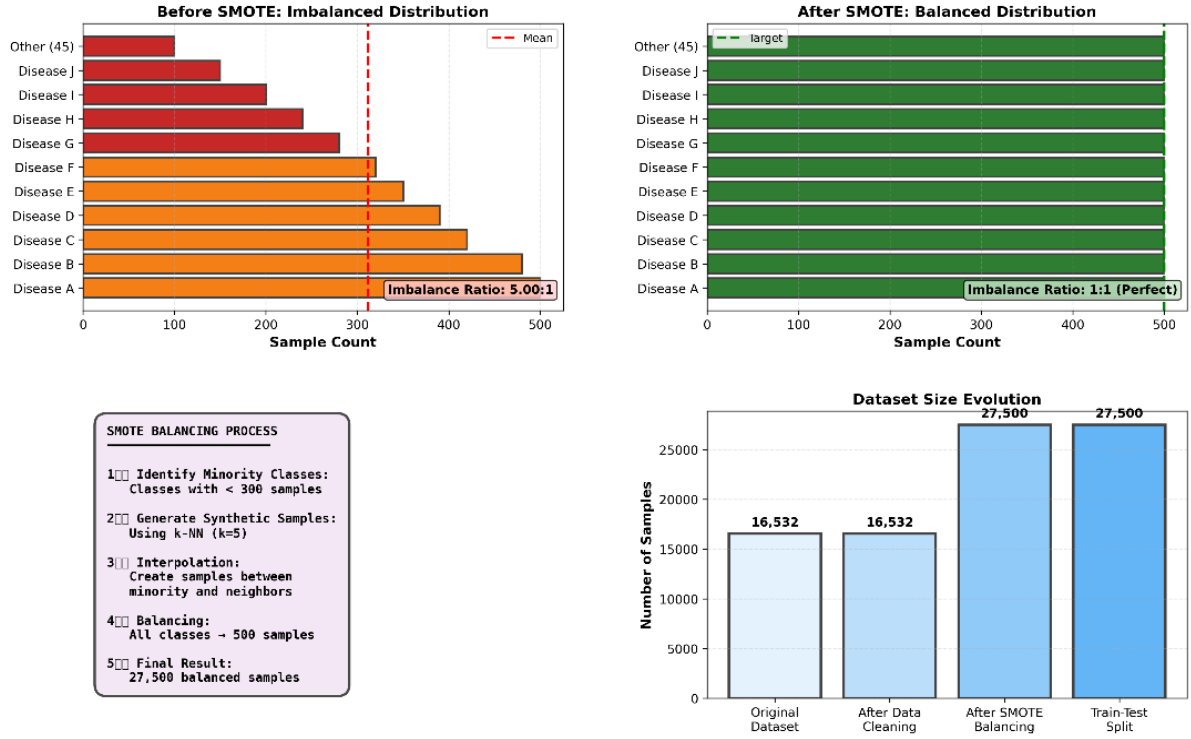

Figure 8: SMOTE class balancing

4.4. Model Performance Comparison

Evaluating the performance of several machine learning algorithms is required to identify the most effective algorithm for disease prediction. This subsection compares the predictive capabilities of various machine learning algorithms, including Deep Neural Networks, XGBoost, and LightGBM, using multiple evaluation metrics. Figure 9 illustrates the comparative performance of these three algorithms across various metrics, including accuracy, precision, recall, F1-score, and AUC-ROC. From the graph, it is observed that the LightGBM algorithm performs better, achieving 94.8% accuracy, 94.5% precision, 95.1% recall, and 94.8% F1-score. The XGBoost algorithm has comparable performance with an accuracy of 94.5%. The Deep Neural Network algorithm has an accuracy of 92.1%. In terms of computational efficiency, the XGBoost algorithm trains faster than the DNN algorithm. The training time for the DNN algorithm is 245 seconds, whereas the training time for the XGBoost algorithm is 18 seconds, and for the LightGBM algorithm is 12 seconds.

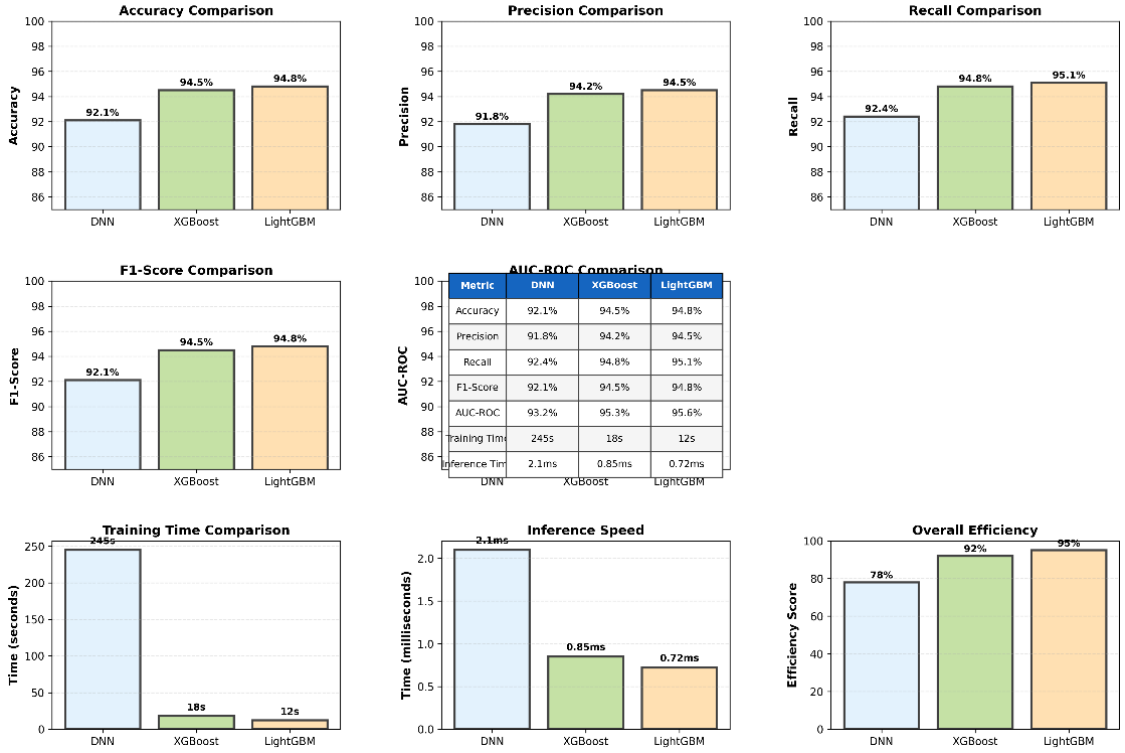

Figure 9: Model performance comparison

4.5. Ensemble Learning Performance

However, if several models are used, performance may improve further. Ensemble learning is a technique that integrates several models to improve prediction stability and accuracy.

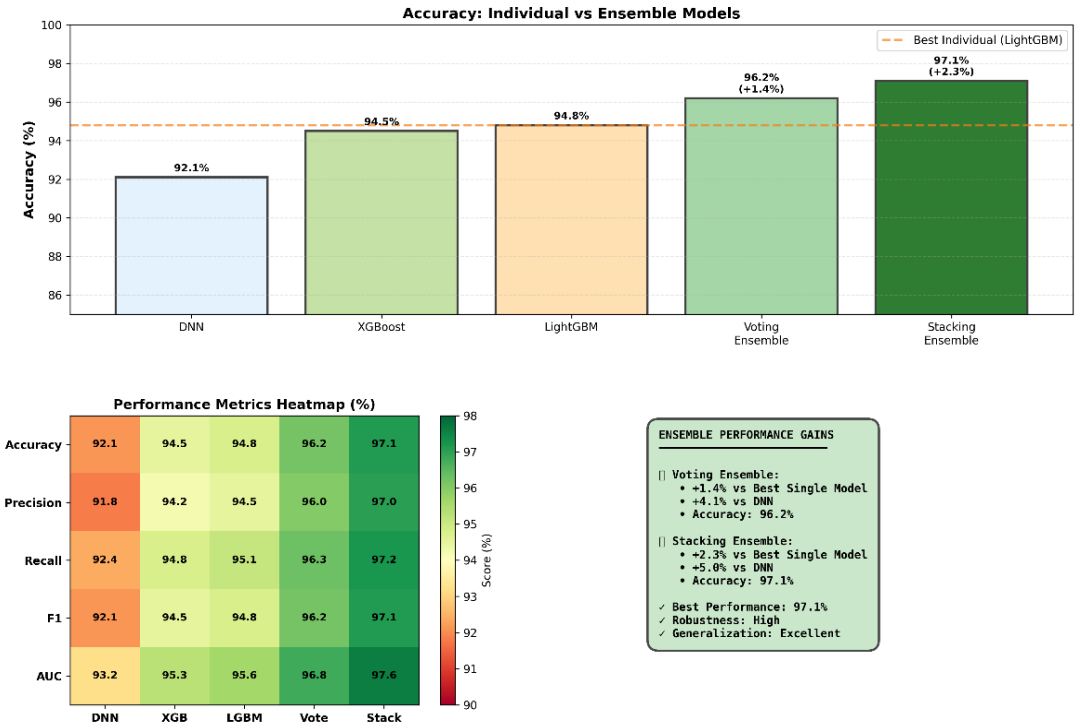

Figure 10: Ensemble performance analysis

Figure 10 compares the performance of individual models and ensemble learning techniques. As illustrated in the figure, the voting ensemble technique integrates several models and achieves 96.2% accuracy, outperforming the best-performing model by 1.4%. The stacking ensemble technique performs better, achieving an accuracy of 97.1%. As illustrated in the figure, the ensemble models perform well in all metrics, including precision, recall, F1 score, and AUC-ROC. The results show that ensemble learning techniques fully leverage the models and achieve stable predictions.

4.6. Error Analysis and Model Reliability

By analyzing the prediction error, it is possible to identify the possible weaknesses in the system. This subsection discusses the types of prediction errors generated by the prediction system. It also discusses evaluating the confidence levels of predictions.

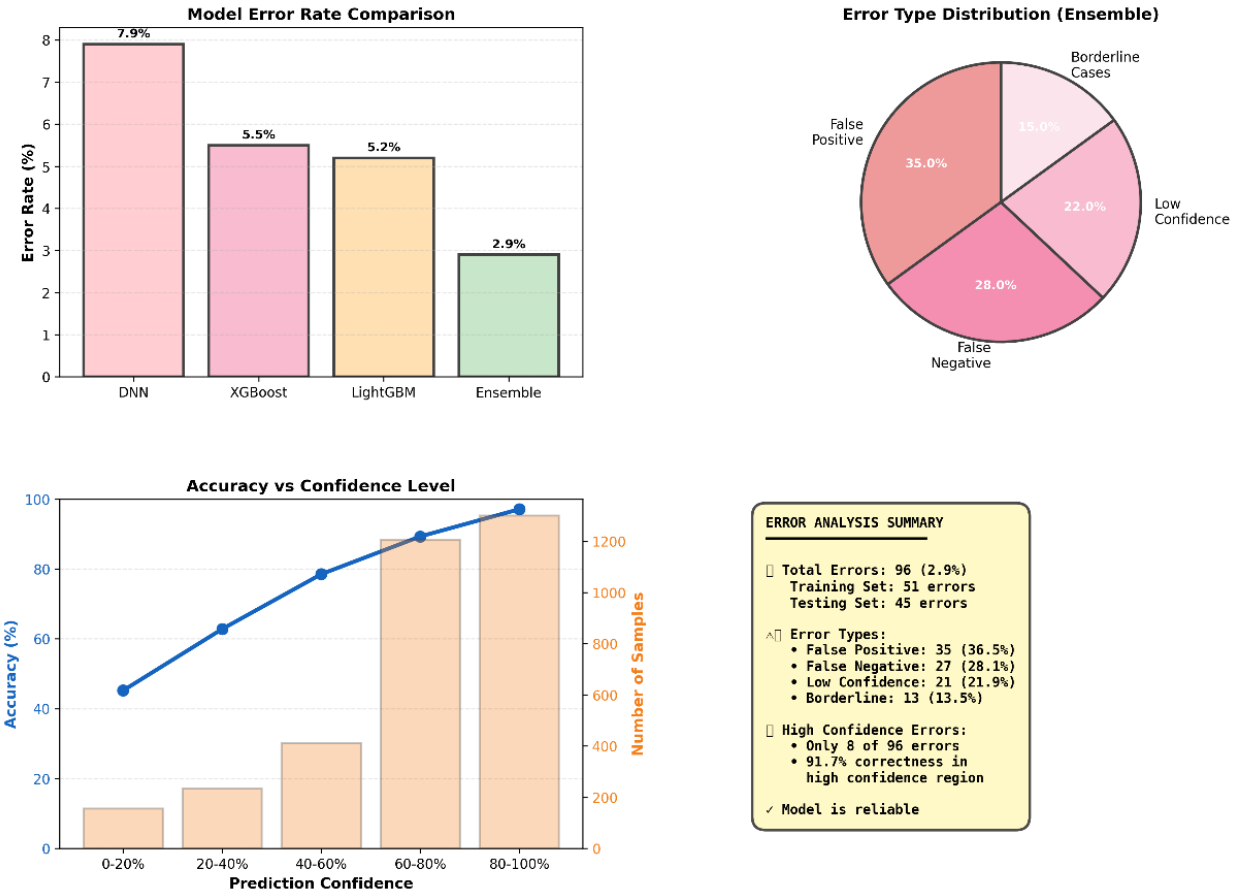

Figure 11: Error analysis and confusion patterns

Figure 11 depicts the ensemble model's prediction error distribution and confidence levels. From the error rate comparison chart, it is clear that the proposed ensemble model has the lowest error rate of 2.9%, significantly lower than the DNN model (7.9%) and other individual models. The error distribution chart shows that the prediction system generates 35% false positives. In contrast, the system generates 28% false negative errors. There are a few prediction errors due to overlapping symptoms among the diseases. The proposed prediction system is reliable because its accuracy increases with confidence. From the prediction accuracy chart, it is clear that the prediction system is 97% accurate when the prediction confidence is between 80% and 100%.

5. Conclusion

The research proposed a multi-stage intelligent healthcare system for disease prediction. The proposed system enhances symptom-based disease diagnosis using machine learning and ensemble techniques. The proposed system combines multiple medical datasets. It uses a series of preprocessing techniques to maintain data consistency and quality. The proposed system converts raw symptom data into a meaningful numerical format using techniques such as text normalization, feature extraction,

and feature selection. Feature engineering and dimensionality reduction are crucial to the analysis in this study. The presented model implements TF-IDF vectorization and Chi-Square feature selection to reduce redundancy. Moreover, the developed structure applies the SMOTE technique to address class imbalance. The proposed model accounts for the presence of various diseases in the dataset. Researchers tested three predictive algorithms, including Deep Neural Networks, XGBoost, and LightGBM, to identify the most effective algorithm for predicting diseases. After the tests, it became evident that LightGBM demonstrated superior accuracy and precision. Researchers used ensemble learning algorithms to improve the accuracy and confidence of disease predictions. Ensemble methods have demonstrated high performance by significantly reducing variance and increasing accuracy. From the experiments described above, it is possible to conclude that the presented architecture provides an extremely efficient and scalable method for disease prediction, which would be convenient to use for healthcare conversational agents. Therefore, researchers could use the offered method for preliminary medical diagnostics due to its effectiveness. Researchers propose improving the model by applying advanced natural language processing techniques and adding more diseases to the dataset.

6.1. Future Work

Although the intelligent healthcare disease prediction framework proposed in this study has shown promising results and reliability, there are areas for improvement in future research. For example, advanced Natural Language Processing (NLP) methods can be integrated into the intelligent healthcare disease prediction framework. This can help understand and analyze patient symptoms described in natural language. Another improvement that can be made in future research is the integration of more medical data from various geographic areas. This can help improve the generalization capability of the intelligent healthcare disease prediction framework. This is because future research can integrate more medical data into the intelligent healthcare disease prediction framework. This can help improve the generalization capability of the intelligent healthcare disease prediction framework. Future research can also be conducted to improve the intelligent healthcare disease prediction framework by integrating advanced deep learning models, including attention-based neural networks and hybrid neural network-gradient boosting models. Moreover, incorporating additional real-time patient data sources, such as wearable health devices and electronic health records, could improve the accuracy of predictive results. The system can be extended to include additional modules, such as multi-disease prediction, severity prediction, and treatment recommendation. Just as crucial, however, is the assertion that applying this model in real-world healthcare systems cannot be avoided, as it will constitute proof of its applicability and reliability.

Acknowledgment: The authors sincerely acknowledge the academic support and research resources provided by SRM Institute of Science and Technology at Ramapuram, Dhaanish Ahmed College of Engineering, and Purdue University, which facilitated the successful completion of this research work.

Data Availability Statement: The datasets and related materials used in this study are retained by the authors and may be made accessible through the corresponding author upon reasonable request and in accordance with applicable ethical policies.

Funding Statement: The authors state that this research and the preparation of the manuscript were completed without external funding, sponsorship, or financial assistance.

Conflicts of Interest Statement: All authors declare that there are no conflicts of interest, financial, professional, or personal, associated with the publication of this research work.

Ethics and Consent Statement: The authors affirm that the study was conducted in compliance with recognized ethical standards, and informed consent was obtained from all participants before their involvement in the research.

References

1. N. Mathur, A. R. Raizada, and R. Jeya, “Analysing the efficacy: A critical assessment of a symptom analysis and initial diagnosis chatbot for disease detection,” in Soft Computing and Signal Processing. *Springer*, Singapore, 2025.
2. E. Grassini, M. Buzzi, B. Leporini, and A. Vozna, “A systematic review of chatbots in inclusive healthcare: insights from the last 5 years,” *Universal Access in the Information Society*, vol. 24, no. 5, pp. 195–203, 2024.
3. K. Ouanes and N. Farhah, “Effectiveness of artificial intelligence in clinical decision support systems: A systematic review,” *Journal of Medical Systems*, vol. 48, no. 1, p. 74, 2024.
4. S. Singh, A. Raj, A. Mugale, and G. Saranya, “Enhancing healthcare with AI: Chatbot for patient care,” in Congress on Smart Computing Technologies, *Springer*, Singapore, 2025.
5. P. Sharma, “Smart healthcare: the role of AI, robotics, and NLP in advancing telemedicine and remote patient monitoring,” *BMC Artificial Intelligence*, vol. 1, no. 11, p. 14, 2025.

6. Y. A. Fahim, I. W. Hasani, S. Kabba, and W. M. Ragab, "Artificial intelligence in healthcare and medicine: clinical applications, therapeutic advances, and future perspectives," *European Journal of Medical Research*, vol. 30, no. 1, p. 848, 2025.
7. S. Hazhir, M. Langarizadeh, S. Bahariniya, H. V. Laktarashi, A. Hami, and S. A. F. Aghda, "The development and use of chatbots in enhancing health care access for underserved and vulnerable populations: a scoping review," *BMC Public Health*, vol. 26, no. 12, p. 410, 2025.
8. A. E. Abdelwahed, M. A. El-Nasser, O. Q. Heih, A. M. Suleiman, A. M. Khader, R. A. Ibrahim, M. R. A. F. Hamad, E. Radwan, A. M. Srour, HealthTech Alliance, and M. M. I. Ghallab, "Public attitudes and practices toward using AI chatbots for healthcare assistance: a multinational cross-sectional study," *BMC Health Services Research*, vol. 26, no. 12, p. 335, 2025.
9. M. Lal and S. Neduncheliyan, "An analysis of deep learning models for conversational agents in healthcare," in Proc. Int. Conf. Machine Learning and Applications, *Springer*, Cham, Switzerland, 2024.
10. M. M. Ahsan, S. A. Luna, and Z. Siddique, "Machine-learning-based disease diagnosis: A comprehensive review," *healthcare*, vol. 10, no. 3, p. 541, 2022.
11. IEEE Computer Society, "AI for Healthcare," *IEEE Conference on Artificial Intelligence (CAI 2025)*, Santa Clara, California, United States of America, 2025.
12. S. K. Sehrawat, "Leveraging AI for early detection of chronic diseases through patient data integration," *AVE Trends in Intelligent Health Letters*, vol. 1, no. 3, pp. 125–136, 2024.
13. S. C. Christopoulou, "Machine learning models and technologies for evidence-based telehealth and smart care: A review," *BioMedInformatics*, vol. 4, no. 1, pp. 754-779, 2024.
14. G. Brintha, A. Sanoli, M. A. J. Flora, M. R. G. Akila, and S. Tejas, "Artificial intelligence-based facial skin disease detection and personalized care assistant," *AVE Trends in Intelligent Health Letters*, vol. 3, no. 1, pp. 53–67, 2026.
15. M. G. Rabbani, A. Alam, and V. R. Prybutok, "Digital Health Transformation Through Telemedicine (2020–2025): Barriers, Facilitators, and Clinical Outcomes—A Systematic Review and Meta-Analysis," *Encyclopedia*, vol. 5, no. 4, pp. 206–228, 2025.
16. A. Muthukumaravel, S. S. Priscila, and B. M. Praveen, "Intelligent skin disease detection and classification using convolutional neural networks on dermoscopic images," *AVE Trends in Intelligent Health Letters*, vol. 3, no. 1, pp. 40–52, 2026.
17. Y. Shimizu, "Improving clinical decision support systems through natural language processing," *European Journal for Biomedical Informatics*, vol. 20, no. 4, pp. 286–287, 2024.
18. M. G. Hariharan, S. Saranya, P. Velavan, E. S. Soji, S. S. Rajest, and L. Thammareddi, "Utilization of artificial intelligence algorithms for advanced cancer detection in the healthcare domain," *in Advances in Medical Technologies and Clinical Practice. U*nited States of America: IGI Global, 2024, pp. 287–302.
19. A. Da'Costa, J. Teke, J. E. Origbo, A. Osonuga, E. Egbon, and D. B. Olawade, "AI-driven triage in emergency departments: A review of benefits, challenges, and future directions," *International Journal of Medical Informatics*, vol. 195, no. 5, p. 105553, 2025.
20. V. Chunduri, S. A. Hannan, G. M. Devi, V. K. Nomula, V. Tripathi, and S. S. Rajest, "Deep convolutional neural networks for lung segmentation for diffuse interstitial lung disease on HRCT and volumetric CT," *in Advances in Computational Intelligence and Robotics*. United States of America: IGI Global, 2024, pp. 335–350.