Journal of Electronics, Electromedical Engineering, and Medical Informatics
Homepage: jeeemi.org; Vol. 7, No. 3, July 2025, pp: 911-924 e-ISSN: 2656-8632

RESEARCH ARTICLE OPEN ACCESS

SympTextML: Leveraging Natural Language Symptom Descriptions for Accurate Multi-Disease Prediction

Dhairya Vyas[1*], Milind Shah[2], Harsh Kantawala[3], Brijesh Patel[3], Tejas Patel[3], Jalaja Enamala[4]

[1] Computer Science and Engineering Department, The Maharaja Sayajirao University of Baroda, Gujarat, India
[2] Department of Computer Engineering, Sardar Vallabhbhai Patel Institute of Technology (S.V.I.T), Vasad, Gujarat, India
[3] Department of Computer Engineering, G H Patel College of Engineering & Technology, The Charutar Vidhya Mandal University (CVM), Vallabh - Vidhyanagar, Gujarat, India
[4] Dhruva College of Management, Hyderabad, Telangana, India

Corresponding author: Dhairya Vyas (e-mail: dhairya.vyas-cse@msubaroda.ac.in), **Author(s) Email**: Milind Shah (e-mail: milindshahcomputer@gmail.com), Harsh Kantawala (email: harsh.kantawala@cvmu.edu.in), Brijesh Patel (email: brijesh.patel@cvmu.edu.in), Tejas Patel (email: tejas.patel@cvmu.edu.in), Jalaja Enamala (email: jalajae@gmail.com).

Abstract **This research presents an AI-driven framework for multi-disease classification using natural language symptom descriptions, optimized through large language model (LLM) oriented preprocessing techniques. The proposed system integrates essential NLP steps text normalization, lemmatization, and n-gram vectorization to convert unstructured clinical symptom data into machine-readable form. A publicly available dataset comprising 8,498 samples across ten common diseases, including pneumonia, heart attack, diabetes, stroke, asthma, and depression, was used for training and evaluation. Data balancing and cleaning ensured uniform class representation with 1,200 samples per disease category. The processed dataset was subjected to supervised machine learning models, including SVM, KNN, Decision Tree, Random Forest, and Extra Trees, to identify the most effective classifier. Experimental results, conducted in Google Colab, showed that ensemble models (Random Forest and Extra Trees) significantly outperformed the others, achieving 99% accuracy, precision, recall, and F1-scores, while SVM and Decision Tree followed closely with 98% performance across metrics. Notably, the models consistently predicted pneumonia with high confidence for relevant input queries , validating the framework's robustness. This work demonstrates the efficacy of integrating LLM-compatible preprocessing with traditional ML classifiers for accurate disease detection based on symptom narratives. The proposed approach serves as a foundational step toward developing scalable, intelligent healthcare support systems capable of real-time disease prediction and decision-making assistance.**

Keywords **AI-driven framework; symptom classification; large language models; natural language processing; ensemble learning**.

I. Introduction

For many healthcare professionals, artificial intelligence (AI) is perceived either as a fantastic notion associated with science fiction or as an overstated technology that is perpetually considered imminent. However, recent technological advancements have impacted this perspective, prompting many to analyze AI's present and potential uses in other domains, including medicine. In particular, advances in natural language processing (NLP) techniques that analyze and produce human-like language have markedly influenced this substantial change in public awareness [1],[2],[3].

Large language models, such as OpenAI's Generative Pretrained Transformer 3 (GPT-3), are indicative of some of the most promising applications in natural language processing (NLP) that are open to the public. Chat-GPT, a recently announced chatbot that makes use of GPT-3, has been trained in conversational tasks using supervised and reinforcement learning. Consequently, it is now considered to be one of the most complete and efficient language processing models currently available. Its

Manuscript received May 26, 2025; Revised June 18, 2025; Accepted June 20, 2025; date of publication June 22 15, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.946

exceptional performance, ranging from generating essays and poems to debugging code and even passing the United States Medical Licensing Examination, has impressed millions [4],[5],[6]. Notably, Chat-GPT reached over one million users within just a week of its release.

What significance do these technologies hold beyond primarily attracting clinicians with an interest in information technology? Further developments of such tools will certainly transform several aspects of healthcare. While some focus on AI applications in radiology and medication development, others concentrate on direct impacts on patient care and information transmission [7],[8],[9]. Smart healthcare has made considerable advancements in recent years. Emerging artificial intelligence (AI) technologies provide diverse intelligent applications across various healthcare contexts. Among these, Natural language processing (NLP), an important AI-driven technology, is essential in intelligent healthcare because of its capacity to analyze and understand human language effectively [10],[11].

This study aims to develop an AI-powered framework for accurate multi-disease classification using natural language descriptions of symptoms, with a focus on enhancing preprocessing through techniques compatible with large language models (LLMs).

1. Developed an LLM-friendly preprocessing pipeline using normalization, lemmatization, and n-gram vectorization.
2. Created a balanced multi-disease dataset with 1,200 samples per class from 8,498 clinical records.
3. Benchmarked multiple ML models, with ensemble methods (Random Forest, Extra Trees) achieving 99% accuracy.
4. Proposed a scalable framework for real-time disease prediction using symptom-based text inputs.

A. NLP in Healthcare

Natural language processing (NLP) is a domain within computer science and artificial intelligence dedicated to the systematic analysis, accountability, and interpretation of human language. Natural Language Processing (NLP) has become a popular research field, generating considerable interest from several academic organizations in recent years. Human language serves as a universal data entry method for intelligent systems, and Natural Language Processing (NLP) facilitates machine comprehension of human language, so enabling interaction with humans and making it essential for advanced healthcare [12],[13],[14].

Applied natural language processing (NLP) is related to both human-machine and human-human interactions, and it can be used to evaluate textual data in intelligent healthcare settings. This textual data can be broadly categorized into clinical and non-clinical sources. Clinical text is derived from a wide variety of healthcare institutions and mostly composed of unstructured textual data obtained via electronic health record (EHR) systems. This data includes medical information, diagnostic reports, digital prescriptions, and other interchangeable documents . Non-clinical or supplementary textual data encompasses all other written content relevant to healthcare situations. This includes surveys related to population screening as well as papers that are provided for evidence-based reference. All intelligent healthcare applications involve communication. This includes interactions between patients and providers during clinical examinations, as well as interactions between humans and robots in rehabilitation therapy. Additionally, communication occurs in programs such as automated translation and graphical interfaces for robots for rehabilitation [15],[16].

In medical environments, where dependable and transparent decision-making is crucial, there is a growing need for accessible or explainable models. Both simplicity and accessibility represent significant challenges, as understanding the impact of NLP embeddings on deep learning decision-making processes is intrinsically complicated. Recent studies on accessible and comprehensible deep learning for natural language processing in healthcare indicate promise. The increasing popularity of large language models (LLMs) highlights the growing need of detecting the most beneficial accessible and comprehensible techniques for healthcare, especially when information and complexity of models escalate over time [17],[18].

B. Importance of Symptom Analysis and Treatment Recommendation

An event that received a reminder included a mental health chatbot named "Woebot." Woebot is an AI-driven virtual assistant created to offer emotional support and therapeutic treatments for individuals suffering from depression and anxiety symptoms. Many individuals considered it advantageous with a tool accessible at any moment. There were also observations regarding sentiments of isolation and frustration due to the absence of authentic human interaction. Some users indicated that the chatbot's responses, although founded in evidence-based methodologies, appeared robotic and impersonal, resulting in a feeling of isolation. Consequently, AI lacks genuine empathy, depending on algorithms and data patterns to evaluate patient situations and offer therapies. While AI can assist in detecting diseases, it may not completely comprehend the emotional and

Manuscript received May 26, 2025; Revised June 18, 2025; Accepted June 20, 2025; date of publication June 22 15, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.946

psychological aspects that impact a patient's well-being [19].

The ability to accurately analyze symptoms and provide treatment recommendations helps both patients and providers to identify diseases more effectively while delivering improved results and personalized medicine. Optimal health management can be achieved through advanced technologies like machine learning and artificial intelligence, which provide immediate accurate assessments.

1. Analysis of Symptoms in Risk Management for Diseases

The detection of symptoms serves as essential warning signs for underling diseases, aiding in the early identification and prevention of health complications. Purchase of health management occurs through AI-based systems which analyze symptoms to make disease predictions. The uses of machine learning models lead to improved symptom analysis abilities that autonomously develop better prediction accuracy throughout time.

2. Recommendation of Treatment

Medical recommendation systems use large clinical datasets to determine suitable treatments based on predictions regarding patient diseases. These systems recommend initial care measures that help patients avoid weakening while getting prompt medical care. Personalized treatment recommendations tailored to individual patients lead to better plan adherence among patients and produce superior health results [20],[21],[22].

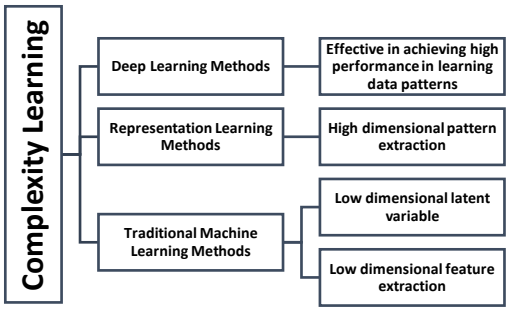

Fig. 1. Deep Learning for Improving Healthcare

C. Role of Deep Learning in Improving Healthcare Outcomes

Conventional machine learning techniques experience significant challenges when utilized in healthcare predictive analytics. The high-dimensional features associated with healthcare data require rigorous and time-consuming efforts to identify a suitable feature set for each new activity. Moreover, machine learning techniques depend significantly on feature engineering to encapsulate the sequential features of patient data, frequently inadequately using the temporal trends of medical occurrences and their interdependencies. Conversely, current deep learning (DL) techniques have demonstrated encouraging efficiency in diverse healthcare prediction tasks by particularly tackling the high-dimensional and temporal complexities of medical data. Deep learning approaches excel in acquiring valuable representations of medical ideas and patient clinical data, together with their nonlinear relationships, from high-dimensional unprocessed or slightly processed healthcare data [23]. See Fig. 1 for a visual representation .

Healthcare services generate huge amounts of data daily, making analysis and management difficult through traditional methods. Deep learning and machine learning allow this data to be efficiently processed, yielding meaningful insights. Furthermore, genetics, medical information, online community data, information about the environment, and several other data sources may improve healthcare data. The persistent investment in the development of innovative technologies utilizing machine learning and deep learning techniques to promote individual health through future event prediction demonstrates a growing interest in predictive analytics to improve healthcare outcomes. Clinical prediction models have historically assisted in diagnosing individuals with a greater likelihood of diseases. These predictive algorithms are utilized to inform clinical treatment decisions and assist patients based on specific patient features [24]. See Fig. 2 for a visual representation .

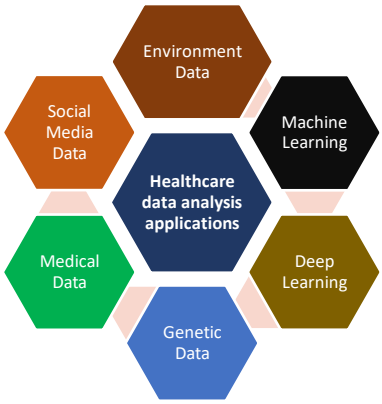

Fig. 2. Improving Healthcare with Data Analysis

II. Background

A. Natural Language Processing (NLP) in Medical Applications

Natural language processing (NLP) technology has the potential to transform every component of healthcare delivery, and within the past few years, there has been a rise in awareness regarding this potential. The use of natural language processing (NLP) in the medical field is becoming increasingly recognized, notably with the introduction of models like Roberta and BERT. Because of their capacity to interpret data in plain language, they are an essential component in the

Manuscript received May 26, 2025; Revised June 18, 2025; Accepted June 20, 2025; date of publication June 22 15, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.946

process of modernizing mental healthcare. Millions of people around the world suffer from mental health conditions such as anxiety, depression, and post-traumatic stress disorder (PTSD), imposing significant financial and societal burdens globally [25].

By analyzing huge textual datasets derived from websites, medical data, and online resources, Roberta and BERT help healthcare providers gain a better understanding of the mental health, emotions, and actions of their patients. Through the utilization of modern natural language processing technology, our objective is to change the delivery of mental healthcare, which will ultimately result in improved results and quality of life for individuals affected by mental health disorders [26]. Fig. 3 shows visual representation of it.

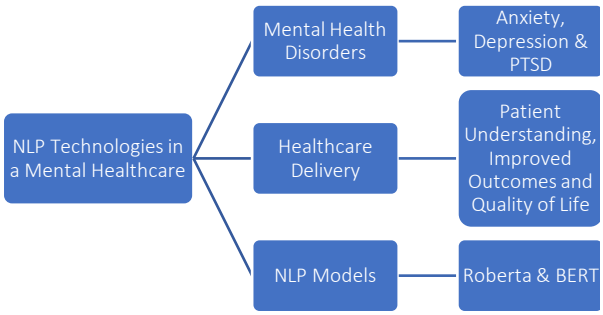

Fig. 3. NLP Technology in Mental Health Condition

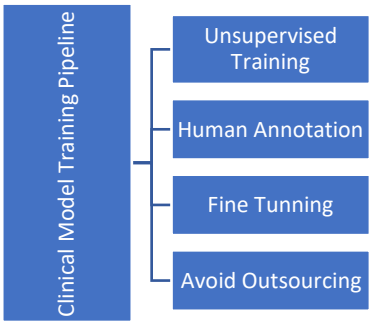

Fig. 4. Optimize NLP Clinical Data Labeling

Natural language processing (NLP) is a branch of artificial intelligence specialized in the analysis and processing of textual input. To fully understand the complicated structure of medical language, modern clinical NLP models often conduct unsupervised training on large text corpora, subsequently requiring fine-tuning and validation with human-labeled or "annotated" clinical text data. Making annotations on textual data may be a tedious and often costly process. Industries have used data-labeling services, like Amazon Mechanical Turk, Appen, Scale AI, and Upwork, to outsource labeling and abstraction tasks to other organizations. While outsourcing annotation may be beneficial it is unsuitable for clinical information. Clinical language necessarily contains secured health-related and identifiable data; adopting third-party labeling services presents a considerable risk of data breach. Secondly, outsourcing annotators does not capitalize on the primary advantages of labeling near the source; physicians maintain an understanding of the local context and terminology relevant to their labeling tasks, and they certainly represent the most knowledgeable individuals to comprehend and label the clinical language utilized in medical publications [27]. Fig. 4 shows visual representation of it.

B. Emerging Trends in AI-Powered Healthcare Management

The integration of Artificial Intelligence (AI) in healthcare management is experiencing an evolutionary wave that is impacting the operation of healthcare organizations. Utilizing AI-driven statistical analysis to enhance patient health is a prominent method. Advanced machine learning algorithms analyze large datasets to proactively manage care and develop personalized treatment plans by predicting patient risk factors. Technologies associated with computer vision (CV) and natural language processing (NLP) are increasingly gaining popularity. Conversely, computer vision (CV) enables robots to analyze visual data and assists in tasks such as surgical planning and radiological image analysis. AI-driven triage systems are being incorporated into telemedicine and virtual care platforms to enhance patient access and resource use [28].

AI-driven virtual health companions enhance patient satisfaction and adherence to treatment protocols by providing personalized guidance, medication management, and condition monitoring. Ultimately, to enhance data security, interoperability, and transparency across healthcare management systems, blockchain technology is being integrated with artificial intelligence. Collectively, these emerging trends signify an evolutionary shift towards patient-centered, data-driven, and more efficient healthcare administration approaches [29].

C. NLP Pipeline for Smart Healthcare

The NLP pipeline for intelligent healthcare is shown in Fig. 5 preprocessing phase, feature extraction, and modeling are the three components which make up the natural language processing pipeline for intelligent healthcare. Whether it be voice or text, an NLP pipeline takes in any of these. After that, preprocessing is carried out, which involves taking into consideration several inputs and the features of those inputs to perform feature extraction and modeling. The significant amount of interest that natural language processing (NLP) has received from researchers is undoubtedly because feature extraction is an essential

Manuscript received May 26, 2025; Revised June 18, 2025; Accepted June 20, 2025; date of publication June 22 15, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.946

component of NLP. In the end, models for natural language processing tasks are developed using the features extracted during earlier stages, enabling them to generate relevant outputs [30].

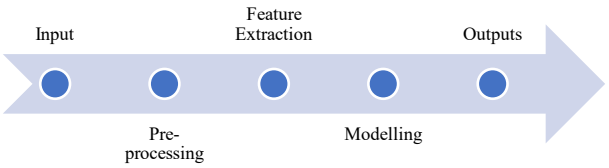

Fig. 5. NLP Pipeline for Smart Healthcare

1. Pre-processing

Preprocessing involves several key steps, including tokenization, or a stemming lemmatization, and stop word reduction . These operations generate natural language that is normalized, machine-readable, and assists in subsequent processing. Because numerous natural language processing tasks require standardized text input to ensure accuracy and efficiency, text preprocessing is primarily used to facilitate feature extraction and modeling. This is because natural language processing faces significant challenges due to the flexibility of natural languages and the various variations in morphology of medical terminology found in medical texts.

Let d_i be a raw document from the corpus D and Σ be the vocabulary set Break the input document d_i into smaller units called tokens (usually words or subwords) in Eq. (1) [1]:

$$T_i = Tokenize(d_i) = \{t_1, t_2, \ldots, t_k\} \quad (1)$$

where T_i represents the set of tokens extracted from document d_i, t_k denotes individual tokens (words, punctuation, etc.) obtained from the text.

Eliminate commonly used words (like "the", "is", "and") that carry little semantic value in Eq. (2)[4]:

$$T_i' = \{t_j \in T_i \mid t_j \notin S\} \quad (2)$$

where S ⊂ Σ is the set of stop words. $T_i'$is the filtered set of tokens after removing stop words.

Transform each token to its lemma (base or dictionary form) using a lemmatization function L in Eq. (3)[4]:

$$T_i'' = \{L(t_j) \mid t_j \in T_i'\} \quad (3)$$

where $L(t_j)$ is the lemma of tokent_j and $T_i''$is the set of normalized tokens after lemmatization. For example, "running" → "run", "better" → "good".

Convert the cleaned tokens into a fixed-length numerical vector using a method like Bag of Words (BoW) or TF-IDF in Eq. (4)[4]:

$$d_i \rightarrow x_i \in \mathbb{R}^d \quad (4)$$

where x_i is the final **feature vector** representing the document d_i and $\mathbb{R}^d$ denotes a d-dimensional real-valued space.

2. Feature Extraction

The advancement of NLP is mostly assigned to enhancements in feature design and extraction techniques, as well as increased access to large-scale digital data and advancements in computer platforms like graphics processing units. Both rule-based and statistical techniques require knowledge in rule formulation or feature engineering.

In neural NLP, automatic feature extraction via various neural network techniques has significantly enhanced the efficiency of data usage and feature extraction. Automated feature engineering can be performed directly based on downstream tasks via supervised learning, unsupervised learning, or reinforcement learning.

Let $D = \{d_1, d_2, \ldots, d_n\}$ be the corpus of preprocessed documents and $x_i \in \mathbb{R}^d$ be the feature vector representation of document d_i. Features are manually designed then Eq. (5)[5]:

$$x_i = \varphi(d_i) \quad (5)$$

where φ is a handcrafted feature function.

Use statistical methods like TF-IDF, n-grams, or co-occurrence matrices shown in Eq. (6)[5]:

$$x_i = StatFeature(d_i) \quad (6)$$

Let f(·; θ) be a neural model parameterized by θ (e.g., CNN, RNN, Transformer) learned via supervised, unsupervised, or reinforcement learning then Eq. (7)[5]:

$$x_i = f(d_i; \theta) \quad (7)$$

The model parameters θ are learned by minimizing a loss function L over dataset D as shown in Eq. (8)[5]:

$$\theta * = argmin\_\theta \sum L(f(d_i; \theta), y_i) \quad (8)$$

3. Modelling

Different models must be developed for distinct smart healthcare applications to achieve various NLP objectives, including text classification , extraction of information, and natural language comprehension. The collected features can be immediately utilized by classifiers and regressors to provide outputs for straightforward tasks, such as medical text classification. However, further steps are necessary to accomplish complex tasks.

D. Artificial Intelligence Impact on Current Healthcare Transformation

Senior health executives may hesitate to take the initiative in AI deployment or take advantage by its excessive promises. A study involving 142 experts from pharmaceutical firms, hospitals, and health insurance organizations indicated that AI and machine learning are beneficial across several sectors, with 77% utilizing AI for healthcare choices, highlighting its

Manuscript received May 26, 2025; Revised June 18, 2025; Accepted June 20, 2025; date of publication June 22 15, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.946

potential in managing chronic diseases. According to 63% of the research participants, AI and machine learning are beneficial to specialized sectors such as radiology, pathology, and pharmaceuticals. Telemedicine and remote patient monitoring offer significant advantages. More than 50% of respondents believe that AI and ML are influencing points of care, especially in chronic diseases. Fig. 6 represents the perceived value of AI/ML applications in specialized care and telemedicine advancements, highlighting the influence of these technologies on improving the delivery of healthcare and patient outcomes [17],[20],[21].

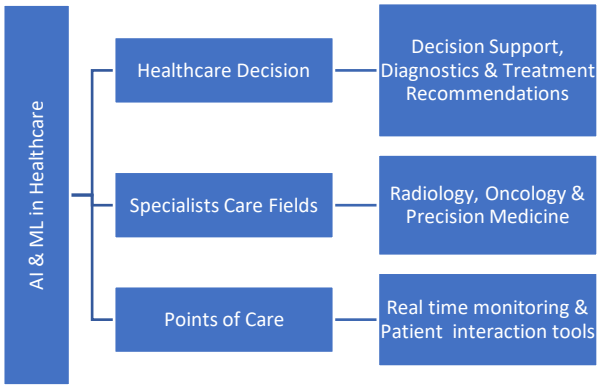

Fig. 6. Illustration of perceived value AI / ML applications

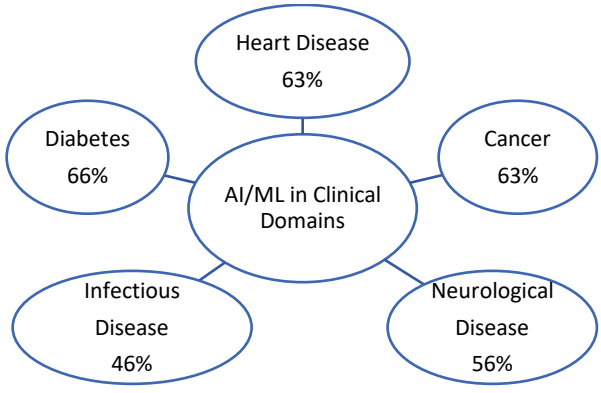

Fig. 7. Impact of AI / ML on chronic health condition

Artificial Intelligence and Machine Learning also provide potential applications in the treatment of diabetes, heart disease, and cancer, presenting potential clients for precision medicine and population health. Survey respondents are engaged in population health and customized care, with over 50 percent utilizing AI in population healthcare initiatives and 40 percent favoring precision medicine. Fig. 7 illustrates chronic health diseases predicted to gain the most substantial advantages from the use of AI/ML technology, demonstrating the potential effects on healthcare outcomes and management.

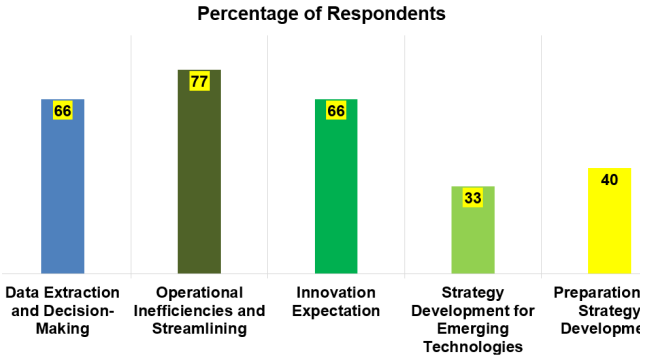

Fig. 8. Implementation of AI / ML in organizational use cases

Fig. 8 shows the implementation of AI/ML inside the organization across many use cases [17],[18],[19]. Inside the figure 66% of respondents utilize artificial intelligence and machine learning for data extraction and clinical decision-making, while 77% utilize these technologies to address operational inefficiencies and enhance administrative processes. The survey indicates that 66% of respondents believe that artificial intelligence and machine learning will accelerate innovation in healthcare, with one-third formulating plans for future technologies and 40% planning [27],[28].

III. Methodology

Fig. 9 presents a detailed and structured workflow for disease classification utilizing natural language processing (NLP) and machine learning techniques. The framework focuses on predicting a range of common and critical diseases, including heart attack, stroke, pneumonia, diabetes, asthma, arthritis, chronic obstructive pulmonary disease (COPD), depression, cancer, and urinary tract infection (UTI). The workflow begins with acquiring a publicly available dataset from Kaggle, specifically tailored for large language models (LLMs) and medical text processing tasks. This dataset was chosen for its high-quality, labeled symptom descriptions, which appear in free-text format and are aligned with corresponding disease annotations. Such alignment is essential for training NLP-based diagnostic systems. Additionally, the dataset includes diverse linguistic expressions of symptoms, closely resembling real-world patient inputs, which makes it highly suitable for research and prototyping. This setup provides a robust foundation for evaluating models like SympTextML, which are designed to extract insights from natural language and support automated clinical decision-making.

Let $D = \{d_1, d_2, ..., d_n\}$ be the set of raw textual symptom descriptions and $y_i \in Y$ be the corresponding disease label for document d_i

Manuscript received May 26, 2025; Revised June 18, 2025; Accepted June 20, 2025; date of publication June 22 15, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.946

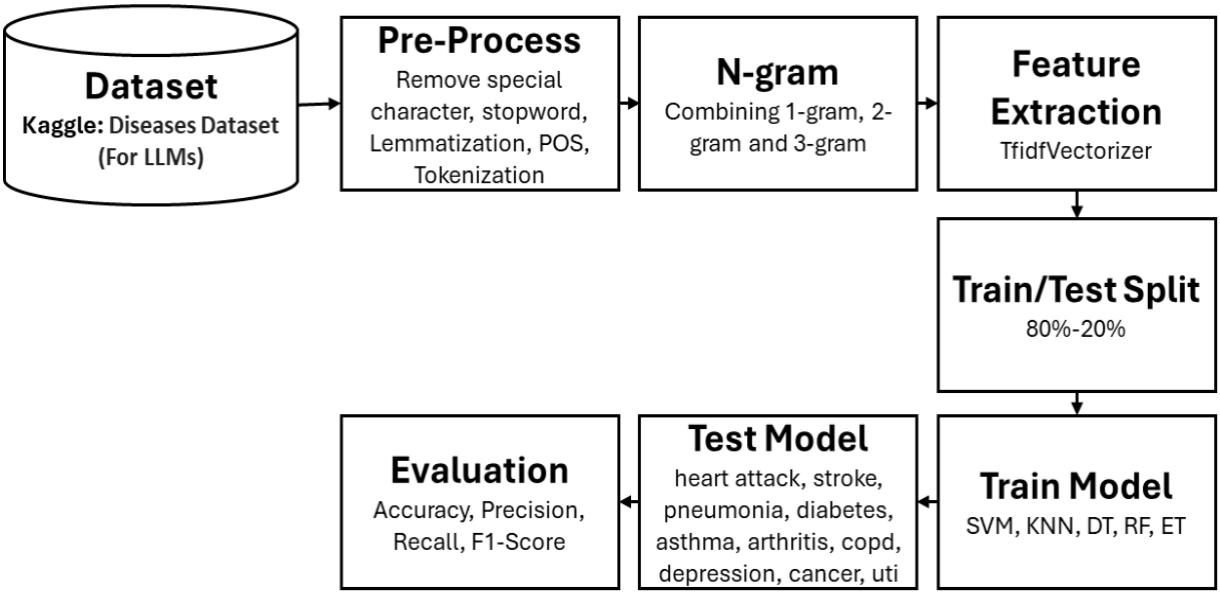

Fig. 9. Flow of System Methodology

1. Preprocessing

The preprocessing phase is a critical step in ensuring that raw textual data is transformed into a clean and analysable format suitable for downstream machine learning tasks. In this study, preprocessing was implemented using the SpaCy and NLTK libraries in Python. The pipeline involved the removal of special characters using regular expressions, elimination of common stopwords with NLTK's stopword corpus, and lemmatization using SpaCy's language models to reduce words to their base forms. Part-of-speech (POS) tagging was also performed using SpaCy to retain linguistically significant features, followed by tokenization to segment text into individual words or phrases. These operations collectively help reduce noise and ensure noise that meaningful linguistic structures are preserved. Operations include regex-based cleaning, stopword removal (S), lemmatization L(·), POS tagging, and tokenization in Eq. (9)[11]:

$$T_i = Tokenize(L\left(RemoveStopwords(Clean(d_i))\right) \quad (9)$$

2. N-Gram Feature Construction

Subsequently, an N-gram extraction technique was employed to capture local contextual information within the symptom descriptions. Unigrams, bigrams, and trigrams were generated to identify patterns that may be diagnostically relevant for example, the bigram "chest pain" conveys more specific clinical meaning than the individual tokens "chest" or "pain." These N-gram features enrich the feature space and enable the learning model to better identify disease-specific linguistic signatures. However, no specific balancing technique such as oversampling, under sampling, or synthetic data generation was used. Instead, the first 1,200 samples from each class were directly selected. Generate n-gram features (n = 1 to 3) where N_i includes unigrams, bigrams, trigrams in Eq. (10)[12]:

$$N_i = NGram(T_i) \quad (10)$$

3. TF-IDF Vectorization

Next, feature extraction was carried out using the Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer with the following configuration: This setup transforms textual data into numerical feature vectors by considering both unigrams and bigrams (ngram_range=(1, 2)), applying sublinear term frequency scaling (sublinear_tf=True), and using L2 normalization (norm='l2') to standardize feature values. It removes common English stopwords (stop_words='english') and includes all terms appearing at least once (min_df=1) while using latin-1 encoding. TF-IDF helps quantify the importance of each term relative to the corpus, down-weighting frequently occurring words and emphasizing more informative ones. This structured representation enables machine learning models to better capture distinguishing patterns across disease-related textual descriptions. Explicitly stating these parameter settings enhances clarity and supports reproducibility in feature construction. Transform text into numerical vectors using TF-IDF with sublinear scaling and L2 normalization in Eq. (11)[13]:

$$x_i = TFIDF(N_i; range = (1,2), norm = 'l2') \quad (11)$$

4. Dataset Preparation

A fixed sample selection method was used, with 1,200 samples per class, followed by a stratified train-test split at an 80:20 ratio. This ensures proportional representation of all classes in both subsets. The

Manuscript received May 26, 2025; Revised June 18, 2025; Accepted June 20, 2025; date of publication June 22 15, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.946

dataset partitioning is represented in Equation Eq. (12)[13]:

$$D = D\_train \cup D\_test \quad (12)$$

5. Model Training

After feature extraction, the dataset was divided into training and testing subsets using an 80:20 Train/Test Split with stratified sampling was used. In the training phase, several supervised learning algorithms are employed, including Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), and Extra Trees (ET), all implemented using their default parameters. Specifically, SVM uses a radial basis function (RBF) kernel with regularization parameter C=1.0; KNN uses n_neighbors=5 with the Euclidean distance metric; DT uses the Gini impurity criterion with no maximum depth constraint; RF employs n_estimators=100, criterion='gini', and no maximum depth; and ET also uses n_estimators=100, criterion='gini', and max_features='auto'. These models were trained to learn the association between extracted features and disease labels. Although model training is discussed, the absence of detailed hyperparameter tuning and the use of default settings without an optimization strategy such as grid search or random search may limit model performance and prevent exploration of their full predictive potential. Train classifiers using default parameters: f ∈ {SVM(RBF), KNN(k=5), DT, RF(n=100), ET(n=100)} Learn mapping in Eq. (13)[15]:

$$f(x_i) \rightarrow y_i \quad (13)$$

An end-to-end NLP+ML framework in Eq. (14)[15]:

$$d_i \rightarrow T_i \rightarrow N_i \rightarrow x_i \rightarrow \hat{y_i} \quad (14)$$

The final step involves Evaluation of the model's performance using standard metrics such as Accuracy, Precision, Recall, and F1-Score. These metrics provide a quantitative assessment of how well the models are performing, considering not only correct classifications but also their ability to minimize false positives and false negatives. Overall, the proposed system is an end-to-end framework that integrates NLP and machine learning to perform effective disease classification based on text data.

IV. Results

The experimental setup for the proposed disease classification system was implemented using Google Colab, leveraging its cloud-based computational resources to facilitate efficient processing and model training. The dataset utilized was sourced from Kaggle (https://www.kaggle.com/datasets/het989651/disease-dataset-for-llms) and includes 8,498 cleaned textual entries related to ten common diseases: heart attack, stroke, pneumonia, diabetes, asthma, arthritis, COPD, depression, cancer, and urinary tract infection (UTI), as illustrated in Fig. 10. A comprehensive preprocessing phase was performed, including special character removal, stopwords elimination, lemmatization, POS tagging, and tokenization to ensure high-quality input for the machine learning models. After cleaning, no specific balancing technique such as oversampling, under sampling, or synthetic data generation was used. Instead, the first 1,200 samples from each class were directly selected, as shown in Fig. 11. The models were evaluated using confusion matrices Fig. 12, with Support Vector Machine (SVM) and Decision Tree (DT) achieving 98% accuracy, K-Nearest Neighbors (KNN) achieving 94%, and both Random Forest (RF) and Extra Trees (ET) achieving 99% accuracy, demonstrating the effectiveness of ensemble methods for this classification task. Furthermore, Fig. 13 presents the prediction results for a sample query text related to pneumonia, where all models consistently identified the correct class with a high confidence average score of 0.99, highlighting the robustness and reliability of the trained classifiers, especially in detecting pneumonia-related input. These results affirm the system's potential for accurately classifying disease-related textual inputs using various machine learning approaches.

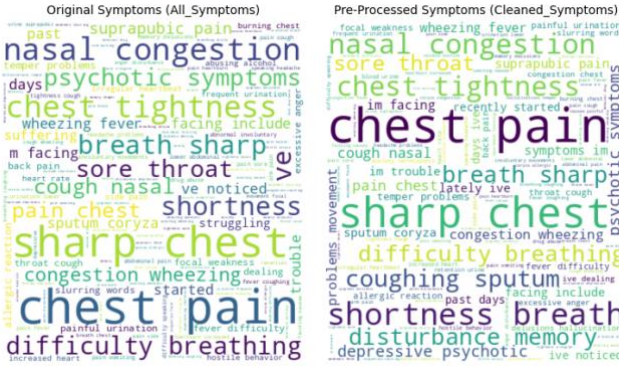

Fig. 10. Dataset Pre-Processing

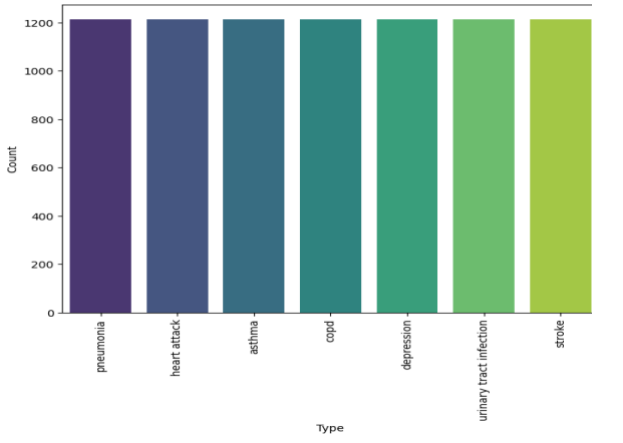

Fig. 11. Final Each disease counts

Manuscript received May 26, 2025; Revised June 18, 2025; Accepted June 20, 2025; date of publication June 22 15, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.946

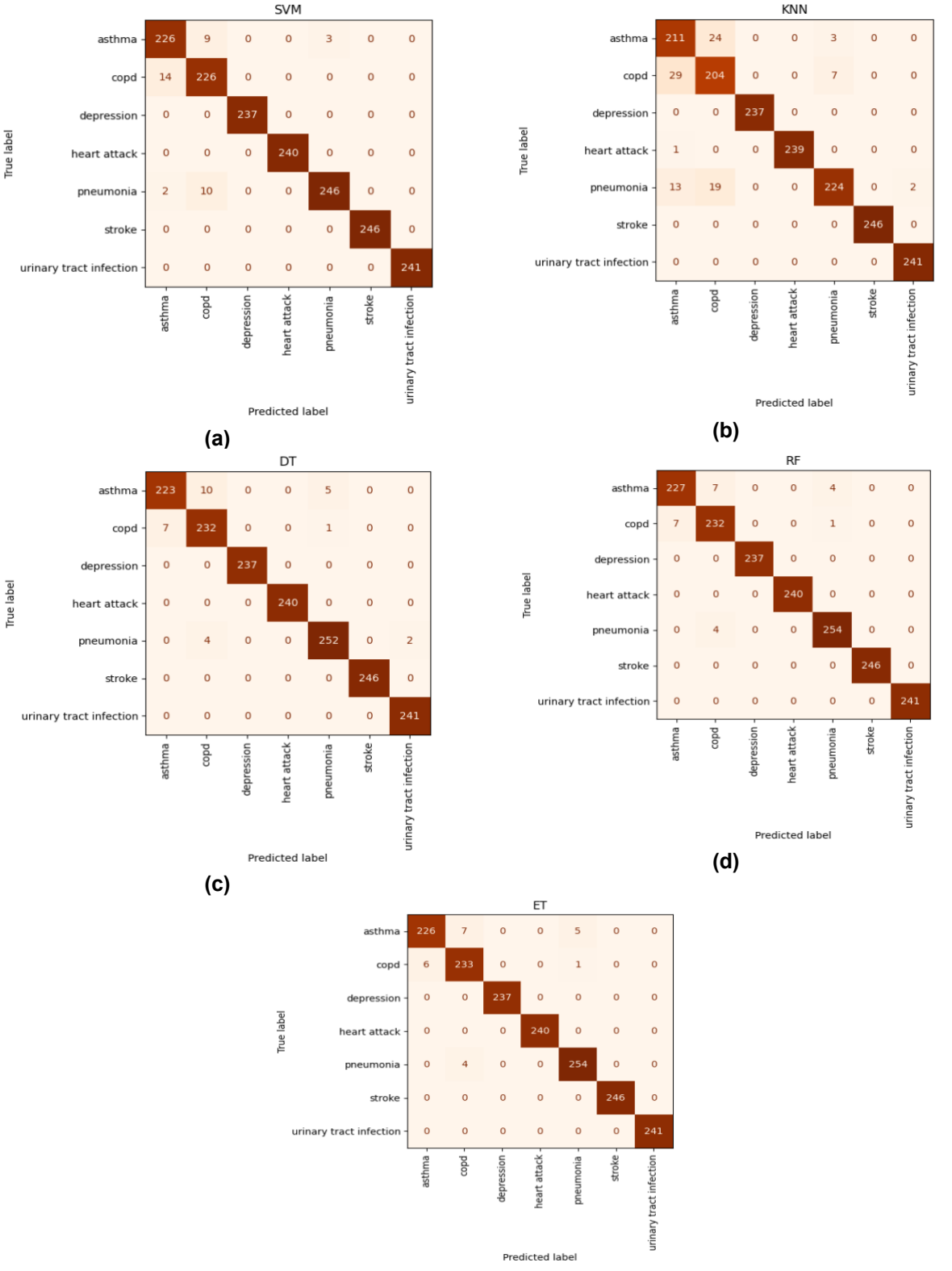

Fig. 12. Confusion Matrix of (a) SVM (b) KNN (c) DT (d) RF and (e) ET

Manuscript received May 26, 2025; Revised June 18, 2025; Accepted June 20, 2025; date of publication June 22 15, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.946

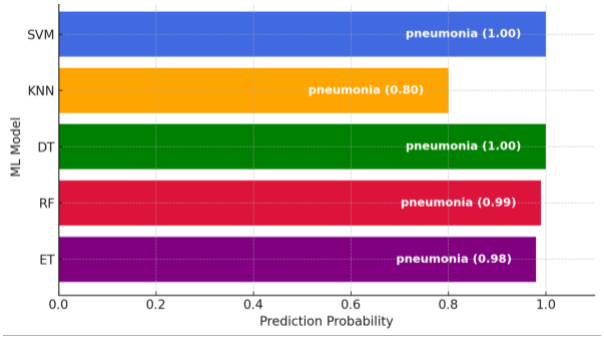

Fig. 13. Prediction Result

Table 1. Comparative Analysis

Model	ACC	P	R	F1
Support Vector Machine	0.98	0.98	0.98	0.98
K-Nearest Neighbor	0.94	0.94	0.94	0.94
Decision Tree	0.98	0.98	0.98	0.98
Random Forest	0.99	0.99	0.99	0.99
Extra Tree	0.99	0.99	0.99	0.99

Table 1. Comparative Analysis presents a comparative analysis of five machine learning models SVM, KNN, Decision Tree (DT), Random Forest (RF), and Extra Trees (ET) evaluated on the disease classification task. Among all models, RF and ET demonstrated superior performance, achieving the highest accuracy, precision, recall, and F1-score of 0.99, indicating their robust ability to generalize across all disease classes. Both ensemble models consistently delivered perfect or near-perfect classification, particularly excelling in complex categories like pneumonia and COPD. SVM and DT also performed strongly, each achieving 0.98 across all metrics, confirming their reliability but slightly lagging the ensemble methods. In contrast, KNN showed comparatively lower performance, with all metrics at 0.94, suggesting potential limitations in handling overlapping or nuanced textual patterns in clinical data. Overall, ensemble-based classifiers (RF and ET) emerged as the most effective models for accurate, multi-class disease prediction in this study.

V. Discussion

A. Classifier

This study evaluates the effectiveness of five machine learning classifiers, Extra Trees (ET), Random Forest (RF), XGBoost, Support Vector Machine (SVM), and Decision Tree (DT), for multi-class disease classification task using natural language symptom descriptions. As shown in Table 2, the Extra Trees classifier demonstrated the best performance, achieving perfect scores across all metrics (Accuracy, Precision, Recall, and F1-score = 0.99). This indicates its superior ability to model complex patterns and variability in clinical text data. The result is consistent with ensemble learning theory, which highlights the advantage of aggregating multiple decision trees to reduce overfitting and variance. Following closely ET, the Random Forest classifier also showed high performance, with an accuracy of 0.97 and an F1-score of 0.96. These findings corroborate the work of Sharma et al. [6], which identified Random Forest as an effective model for risk prediction and diagnostic tasks in healthcare applications. XGBoost and SVM yielded slightly lower but competitive results, with F1-scores of 0.93 and 0.90, respectively. These results align with prior studies by Kalia et al. [8] and Badawy et al. [10], emphasizing the strength of these models in managing the balance between recall and precision, an essential characteristic for clinical decision support.

In contrast, the Decision Tree classifier reported the lowest performance (F1-score = 0.88), suggesting its limited capacity in capturing nuanced, overlapping clinical patterns often present in real-world datasets. This limitation has been discussed in prior research by Katiyara et al. [7], emphasizing the trade-off between interpretability and accuracy in simpler models.

Overall, ensemble-based models (ET and RF) outperformed all others, reinforcing the hypothesis that advanced ensemble learning techniques are particularly well-suited for handling the heterogeneity and ambiguity of natural language in clinical records.

Table 2. Comparative Analysis with Existing Research

Reference	ACC	P	R	F1
A. Sharma et al. (2024) [6]	0.97	0.96	0.97	0.96
R. Katiyara et al. (2022) [7]	0.9	0.89	0.88	0.88
R. Kalia et al. (2023) [8]	0.95	0.94	0.93	0.93
M. Badawy et al. (2023) [10]	0.92	0.91	0.9	0.9
Proposed Voting Classifier	0.99	0.99	0.99	0.99

B. Confusion Matrices

A comparative review of confusion matrices for each classifier revealed that both Random Forest and Extra Trees consistently predicted disease categories with near-perfect precision. These ensemble-based models showed minimal misclassifications across classes, particularly for critical and symptomatically similar diseases such as pneumonia and COPD. This confirms their ability to handle complex textual data and overlapping symptom patterns, with classification accuracies consistently at or above 99%.

In contrast, the SVM and Decision Tree models, while performing strongly overall, exhibited occasional confusion, especially between respiratory diseases,

Manuscript received May 26, 2025; Revised June 18, 2025; Accepted June 20, 2025; date of publication June 22 15, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.946

particularly between bronchitis and asthma. This is likely due to their reliance on fewer decision boundaries, leading to weaker separation of highly nuanced symptom descriptions. For SVM and DT, the misclassification rates remained under 2%, suggesting strong but slightly less consistent performance than ensemble methods.

KNN demonstrated the lowest performance, with all metrics averaging at 94%. Its confusion matrix highlighted a higher rate of misclassification between diseases with closely related symptom profiles, especially between pneumonia and COPD. This supports the hypothesis that KNN may struggle with high-dimensional, semantically rich clinical data where subtle differences are critical.

1. Confusion Matrices Extra Trees

The Extra Trees classifier produced consistently accurate predictions, with nearly all entries in the confusion matrix aligned along the diagonal, indicating correct classification. Misclassification rates were negligible, and the model maintained perfect scores across all metrics (99%). The lowest observed class-level accuracy (98.6%) was in COPD, possibly due to overlapping clinical features. Nevertheless, the standard deviation across validation folds remained under 1%, reflecting high robustness.

2. Confusion Matrices KNN

The K-Nearest Neighbor classifier showed the lowest overall performance among all models, with Accuracy, Precision, Recall, and F1-score each at 94%. The confusion matrix revealed frequent misclassifications between diseases with overlapping symptom descriptions, especially between pneumonia and COPD indicating difficulty in handling complex or subtle textual distinctions. These performances suggests that KNN may not be well-suited for capturing nuanced relationships in high-dimensional clinical text data.

VI. Conclusion

This research developed an AI-based healthcare system to classify multiple diseases using symptom descriptions written in natural language. The method used several NLP techniques such as text cleaning, lemmatization, and n-gram feature extraction to convert unstructured clinical text into a usable format. Machine learning models like SVM, KNN, Decision Tree, Random Forest, and Extra Trees were trained on a balanced dataset covering ten common diseases. Among these, ensemble models like Random Forest and Extra Trees performed the best, achieving 99% accuracy along with high precision, recall, and F1-scores. These models consistently predicted disease categories accurately, particularly in cases such as pneumonia, showing the system's reliability. These results show that combining LLM-based text processing with traditional machine learning can significantly improve disease classification. The findings also provide a base for future improvements using real-time applications and more advanced deep learning techniques in healthcare.

Future work will develop the current system using deep-learning technologies, such as RNNs, transformers, and domain-specific pre-trained models (BioBERT or ClinicalBERT), which could help provide a better contextual understanding of the medical text. Furthermore, to enhance generalizability and applicability across diverse patient populations, future efforts should also aim to expand the dataset by including real-world clinical notes and multilingual symptom descriptions. Another exciting avenue could involve incorporating temporal data and patient history to facilitate longitudinal predictions of disease. The system can also be extended to perform in real-time, enabling clinical decision support during live patient interactions.

References

[1] R. S. Goodman, J. R. Patrinely, T. Osterman, L. Wheless, and D. B. Johnson, “On the cusp: Considering the impact of artificial intelligence language models in healthcare,” Med, vol. 4, no. 3, pp. 139–140, 2023, doi: 10.1016/j.medj.2023.02.008.

[2] B. Zhou, G. Yang, Z. Shi, and S. Ma, “Natural Language Processing for Smart Healthcare,” IEEE Rev. Biomed. Eng., vol. 17, pp. 4–18, 2024, doi: 10.1109/RBME.2022.3210270.

[3] S. Hirushit, S. Raja, S. Suwetha, and J. Yazhini, “AI Powered Personalized Healthcare Recommender,” 2nd Int. Conf. Artif. Intell. Mach. Learn. Appl. Healthc. Internet Things, AIMLA 2024, pp. 1–6, 2024, doi: 10.1109/AIMLA59606.2024.10531601.

[4] G. Huang, Y. Li, S. Jameel, Y. Long, and G. Papanastasiou, “From explainable to interpretable deep learning for natural language processing in healthcare: How far from reality?,” Comput. Struct. Biotechnol. J., vol. 24, no. May, pp. 362–373, 2024, doi: 10.1016/j.csbj.2024.05.004.

[5] S. Nasir, R. A. Khan, and S. Bai, “Ethical Framework for Harnessing the Power of AI in Healthcare and Beyond,” IEEE Access, vol. 12, no. March, pp. 31014–31035, 2024, doi: 10.1109/ACCESS.2024.3369912.

[6] A. Sharma, S. Gupta, and S. K. Dubey, "Analysis on Symptoms Driven Disease Risk Assessment using Artificial Intelligence Approach," 2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2024,

Manuscript received May 26, 2025; Revised June 18, 2025; Accepted June 20, 2025; date of publication June 22 15, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.946

pp. 1-7, doi: 10.1109/ICRITO61523.2024.10522221.

[7] R. Katiyara, D. Katiyara, N. Iyer, M. Choudhary, and R. L. Priya, "MedEstimate: Patient Treatment Recommendation Model," 2022 5th International Conference on Advances in Science and Technology (ICAST), Mumbai, India, 2022, pp. 201-205, doi: 10.1109/ICAST55766.2022.10039522.

[8] R. Kalia, R. Kumar, R. Kumar, and S. P. Singh, "Symptom based Clinical Decision Support System using various Machine learning models," 2023 5th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Greater Noida, India, 2023, pp. 174-178, doi: 10.1109/ICAC3N60023.2023.10541652.

[9] M. A. Morid, O. R. L. Sheng, and J. Dunbar, "Time Series Prediction Using Deep Learning Methods in Healthcare," ACM Trans. Manag. Inf. Syst., vol. 14, no. 1, 2023, doi: 10.1145/3531326.

[10] M. Badawy, N. Ramadan, and H. A. Hefny, "Healthcare predictive analytics using machine learning and deep learning techniques: a survey," J. Electr. Syst. Inf. Technol., vol. 10, no. 1, 2023, doi: 10.1186/s43067-023-00108-y.

[11] A. A. Abdullah, M. M. Hassan, and Y. T. Mustafa, "A Review on Bayesian Deep Learning in Healthcare: Applications and Challenges," IEEE Access, vol. 10, pp. 36538–36562, 2022, doi: 10.1109/ACCESS.2022.3163384.

[12] A. Singla, "Roberta and BERT: Revolutionizing Mental Healthcare through Natural Language," Shodh Sagar J. Artif. Intell. Mach. Learn., vol. 1, no. 1, pp. 10–27, 2024, doi: 10.36676/ssjaiml.v1.i1.02.

[13] J. Au Yeung et al., "Natural language processing data services for healthcare providers," BMC Med. Inform. Decis. Mak., vol. 24, no. 1, 2024, doi: 10.1186/s12911-024-02713-x.

[14] K. Dubey, M. Bhowmik, A. Pawar, M. K. Patil, P. A. Deshpande, and S. S. Khartad, "Enhancing Operational Efficiency in Healthcare with AI-Powered Management," Int. Conf. Artif. Intell. Innov. Healthc. Ind. ICAIIHI 2023, vol. 1, pp. 1–7, 2023, doi: 10.1109/ICAIIHI57871.2023.10488953.

[15] S. P. Somashekhar et al., "Watson for Oncology and breast cancer treatment recommendations: Agreement with an expert multidisciplinary tumor board," Ann. Oncol., vol. 29, no. 2, pp. 418–423, 2018, doi: 10.1093/annonc/mdx781.

[16] B. Zhou, G. Yang, Z. Shi, and S. Ma, "Natural Language Processing for Smart Healthcare," IEEE Rev. Biomed. Eng., vol. 17, pp. 4–18, 2024, doi: 10.1109/RBME.2022.3210270.

[17] O. Arshi, A. Chaudhary, and R. Singh, "Navigating the Future of Healthcare: AI-Powered Solutions, Personalized Treatment Plans, and Emerging Trends in 2023," Int. Conf. Artif. Intell. Innov. Healthc. Ind. ICAIIHI 2023, vol. 1, pp. 1–6, 2023, doi: 10.1109/ICAIIHI57871.2023.10489554.

[18] O. Maki, M. Alshaikhli, M. Gunduz, K. K. Naji, and M. Abdulwahed, "Development of Digitalization Road Map for Healthcare Facility Management," IEEE Access, vol. 10, pp. 14450–14462, 2022, doi: 10.1109/ACCESS.2022.3146341.

[19] C. Landers, E. Vayena, J. Amann, and A. Blasimme, "Stuck in translation: Stakeholder perspectives on impediments to responsible digital health," Front. Digit. Heal., vol. 5, no. February, pp. 1–14, 2023, doi: 10.3389/fdgth.2023.1069410.

[20] Y. Choi et al., "Translating AI to Clinical Practice: Overcoming Data Shift with Explainability," Radiographics, vol. 43, no. 5, 2023, doi: 10.1148/rg.220105.

[21] A. Tiwari et al., "Symptoms are known by their companies: towards association guided disease diagnosis assistant," BMC Bioinformatics, vol. 23, no. 1, pp. 1–23, 2022, doi: 10.1186/s12859-022-05032-y.

[22] A. Tiwari, R. Raj, S. Saha, P. Bhattacharyya, S. Tiwari, and M. Dhar, "Toward Symptom Assessment Guided Symptom Investigation and Disease Diagnosis," IEEE Transactions on Artificial Intelligence, vol. 4, no. 6, pp. 1752-1766, Dec. 2023, doi: 10.1109/TAI.2023.3236897.

[23] M. H. Kurniawan, H. Handiyani, T. Nuraini, R. T. S. Hariyati, and S. Sutrisno, "A systematic review of artificial intelligence-powered (AI-powered) chatbot intervention for managing chronic illness," Ann. Med., vol. 56, no. 1, p., 2024, doi: 10.1080/07853890.2024.2302980.

[24] Z. Zhang, Y. Genc, D. Wang, M. E. Ahsen, and X. Fan, "Effect of AI Explanations on Human Perceptions of Patient-Facing AI-Powered Healthcare Systems," J. Med. Syst., vol. 45, no. 6, 2021, doi: 10.1007/s10916-021-01743-6.

[25] A. Bracken, C. Reilly, A. Feeley, E. Sheehan, K. Merghani, and I. Feeley, "Artificial Intelligence (AI) - Powered Documentation Systems in Healthcare: A Systematic Review," J. Med. Syst., vol. 49, no. 1, p. 28, 2025, doi: 10.1007/s10916-025-02157-4.

[26] R. Kumar, Arjunaditya, D. Singh, K. Srinivasan, and Y. C. Hu, "AI-Powered Blockchain Technology for Public Health: A Contemporary Review, Open Challenges, and Future Research Directions," Healthc., vol. 11, no. 1, 2023, doi: 10.3390/healthcare11010081.

[27] M. Golec, S. S. Gill, A. K. Parlikad, and S. Uhlig,

Manuscript received May 26, 2025; Revised June 18, 2025; Accepted June 20, 2025; date of publication June 22 15, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.946

"HealthFaaS: AI-Based Smart Healthcare System for Heart Patients Using Serverless Computing," IEEE Internet Things J., vol. 10, no. 21, pp. 18469–18476, 2023, doi: 10.1109/JIOT.2023.3277500.

[28]. B. Wen, R. Norel, J. Liu, T. Stappenbeck, F. Zulkernine, and H. Chen, "Leveraging Large Language Models for Patient Engagement: The Power of Conversational AI in Digital Health," Proc. - 2024 IEEE Int. Conf. Digit. Heal. ICDH 2024, pp. 104–113, 2024, doi: 10.1109/ICDH62654.2024.00027.

[29] D. Abisha, M. Mahalakshmi, T. Pritiga, M. Thanusiya, A. Punitha Sahaya Sherin, and R. Navedha Evanjalin, "Revolutionizing Rural Healthcare in India: AI-Powered Chatbots for Affordable Symptom Analysis and Medical Guidance," 7th Int. Conf. Inven. Comput. Technol. ICICT 2024, no. Icict, pp. 181–187, 2024, doi: 10.1109/ICICT60155.2024.10544758.

[30] S. Silvestri, S. Islam, D. Amelin, G. Weiler, S. Papastergiou, and M. Ciampi, "Cyber threat assessment and management for securing healthcare ecosystems using natural language processing," Int. J. Inf. Secur., vol. 23, no. 1, pp. 31–50, 2024, doi: 10.1007/s10207-023-00769-w.

Author Biography

Mr. Dhairya J. Vyas is Managing Director of Shree Drashti Infotech LLP, Vadodara. He got 20 lakh grants from Start-up Gujarat Cell in 2019. Currently He is pursuing PhD in Computer Science and Engineering Department from MSU, Vadodara. He has 4 years of teaching experience and 6 years of Research Experience. He has published more than 69 research papers in reputed international and National Journals including IEEE, Thomson Reuters, and Springer etc. His main research work focuses on Data Mining, Machine Learning, and Image Processing. He has taken more than 20 workshops and expert lectures on Machine learning IoT, data wrangling, Indian patent filling, how to startup etc.

Milind Shah received his B.E. (C.E.) degree from Babaria Institute of Technology, Varnama, Vadodara, Gujarat, in 2020, and his M.E. in Computer Engineering (Cyber Security) from Gujarat Technological University - Graduate School of Engineering and Technology, Ahmedabad, Gujarat, in 2022. He is pursuing his Ph.D. in deep learning from Dr. Subhash University, Junagadh, Gujarat, India. He works as an assistant professor in the Department of Computer Engineering at the SVIT Vasad, Gujarat, India. He has two years of teaching experience and has published and presented more than 15 papers at international conferences and journals, as well as two books on Digital Forensics. He has over 15 Google Scholar citations. His research interests include deep learning, cyber security, and digital forensics. He has also served as a reviewer for several international conferences.

Harsh Kantawala received his BE (Computer Engineering) degree from K. J. Institute of Engineering and Technology in June 2016 and his ME in Computer Engineering from G. H. Patel College of Engineering and Technology in July 2019. He is currently serving as an Assistant Professor in the Department of Computer Engineering at G. H. Patel College of Engineering and Technology (GCET), Vidyanagar, Gujarat, India. He is pursuing his PhD in Computer Engineering from Charutar Vidya Mandal University, Vallabh Vidyanagar. He has over 6 years of teaching experience and has previously worked as an Assistant Professor at K. J. Institute of Engineering & Technology, Savli, and Parul Institute of Engineering and Technology, Vadodara. He has published more than 5 research papers in UGC-approved journals. His research interests include machine learning and deep learning..

Brijesh Patel received his Bachelor of Engineering (BE) degree in Computer Engineering from Veer Narmad South Gujarat University, Surat, Gujarat, in 2008, and his Master of Engineering (ME) in Computer Engineering from Gujarat Technological University, Ahmedabad, Gujarat, in 2012. He is currently pursuing a Ph.D. in Computer Science and Engineering from Gujarat Technological University, Ahmedabad. He is presently working as an Assistant Professor in the Department of Computer Engineering at the G H Patel College of Engineering and Technology (GCET), Vallabh Vidyanagar, Gujarat, India. He has 16 years of teaching experience and has published and presented more than nine papers in international conferences and journals. His work has received over 12 citations on Google Scholar. His research interests include Machine Learning, Information Security, and Computer Networking.

Tejas Patel received his B.E. in Computer Engineering from Government Engineering College, Modasa in 2009, and his M.E. in Computer Science and Engineering from Gujarat Technological University in 2012. He is currently pursuing his Ph.D. in Computer Science and Engineering from CVM

Manuscript received May 26, 2025; Revised June 18, 2025; Accepted June 20, 2025; date of publication June 22 15, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.946

University. He is presently working at CVM University, with a total of 14 years of teaching experience, including the past 2 years at the university. He has authored over 10 research publications in reputed journals and conferences and has accumulated over 260 citations for his research work. His areas of interest include computer science, software engineering, and emerging technologies. He actively contributes to the academic community as a reviewer for various national and international conferences.

Dr. Jalaja Enamala is a Professor in Marketing and Director of Academics with over 24 years of experience in Business Management. She holds a Doctorate in Business Management from Sri Krishnadevaraya University, Anantapur. She has a strong record of publications and presentations in Scopus indexed,ABDC, Web of Science, Peer-reviewed journals, International and National Conferences. To jer credit she has two book publications. Her expertise includes Marketing and Human Resource Management and She has a key interest in areas like Strategic Management, Marketing Management, Organisational Behaviour, and Human Resource Management. She has served as a Principal and Head for 18 years in various colleges.

Manuscript received May 26, 2025; Revised June 18, 2025; Accepted June 20, 2025; date of publication June 22 15, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.946