

Research Digest on Engineering Management and Social Innovations

Available on: <http://rdems.in>

DOI: 10.46647/icetetas094

Volume: 02 Issue: 04, April, 2026

p-ISSN: 3117-4418

e-ISSN: 3117-4426

NLP-Based Multi-Class Disease Classification Using Patient Symptom Text

P. Nikhitha¹, G. Deena Dhanya²

¹Asst Professor at Department of CSE(AI-ML), Andhra Loyola Institute of Engineering & Technology, Vij-8

²Pursuing B. Tech at Department of CSE(AI-ML), Andhra Loyola Institute of Engineering & Technology, Vij-8.

ABSTRACT

Healthcare systems produce large amounts of unstructured patient data in the form of patient symptoms and notes, which pose a challenge in the efficient diagnosis of patients. The current paper aims to develop a model of disease classification using NLP techniques on a synthetic dataset of 25,000 patient records and 30 different categories of diseases. The proposed model applies a text pre-processing technique to classify the diseases using three different classification algorithms to determine the efficiency of the model in terms of accuracy. The text pre-processing technique employed in the proposed model is Bag of Words and Term Frequency-Inverse Document Frequency. The three different classification algorithms employed in the proposed model are Multinomial Naive Bayes, Logistic Regression, and Support Vector Machines. The classification algorithm that achieved the highest accuracy of 94.2%, precision of 93.8%, recall of 93.5%, and F1-score of 93.6% in the proposed model is Support Vector Machines. A simple interface has been employed to fetch the symptoms of the patient to classify the disease by probability-based classification. The proposed model has proved that NLP can be employed as a viable and efficient tool for decision-making.

KEYWORDS

Natural Language Processing(NLP), Disease Classification, Symptom-Based Prediction, TF-IDF, Machine Learning, Multi-Class Classification, Clinical Decision Support System, Healthcare Analytics, Text Mining, Support Vector Machine(SVM).

1. INTRODUCTION

In the healthcare sector, there is a lot of raw or unstructured data about patients in terms of symptoms, clinical notes and eHealth records. To make the right decisions about a patient's diagnosis and make it quickly, you need to be able to effectively analyse these different types of data. Homogeneous, ambiguous free-text symptom descriptions frequently take up a significant amount of time to convert into usable data and are often inconsistent in their interpretation.

1.1 Background

The healthcare industry is data-driven in the modern world. Electronic Health Records, notes, discharge summaries, patient symptoms, etc., form a large unstructured database of health care information. It is one of the biggest natural language databases in the world. According to a study by the World Health Organisation, "more than 80% of the relevant information can be found in unstructured text." The information has to be utilised as efficiently as possible by developing systems that can effectively comprehend natural-language data in the most precise and accurate manner.



Research Digest on Engineering Management and Social Innovations

Available on: <http://rdems.in>

DOI: 10.46647/icetetas094

Volume: 02 Issue: 04, April, 2026

p-ISSN: 3117-4418

e-ISSN: 3117-4426

Natural Language Processing has been identified as the core computer science discipline to address the challenge of unstructured clinical texts in the context of the health care industry. The development in the field of vectorising unstructured texts, classification of texts, and the advent of deep learning have greatly improved the state of the art in the field of medical text mining, symptom mining, and computer-aided diagnosis.

1.2 Problem Definition

The interpretation of patient symptom descriptions by clinicians is a labour-intensive task that is time-consuming and not scalable to the growing number of patients. The development of rule-based expert systems requires the development of large sets of rules that do not generalise to linguistic variations in the patient's symptoms and the development of new symptomologies. The need of the hour is the development of a data-driven NLP-based classification system that can classify patient symptom descriptions into probable diseases as accurately and rapidly as possible.

1.3 Motivation

The motivation for this research comes from three important aspects that need to be addressed in the form of one integrated solution. The first is the need for reduced time in diagnosis in resource-constrained medical settings, which can help reduce avoidable patient morbidity and mortality. The second is the need for developing a machine learning-based system for the classification of symptoms, which can be used for consumer-facing health apps with low computational requirements, meeting the requirements of commodity hardware. The third is the need for developing a system for the classification of symptoms, which can be achieved by employing NLP techniques such as TF-IDF and Linear Classifiers, which can be sufficiently transparent for meeting the requirements of audit trails that are increasingly being mandated by medical regulatory bodies.

1.4 Contribution

The contributions of this research can be listed as follows:

1. Development of a data-driven NLP-powered system for the classification of symptom text, which can be used for the classification of diseases into 30 different diseases.
2. Comparative analysis of Multinomial Naive Bayes, Logistic Regression, Support Machines, with a synthetic large-scale healthcare text data set.
3. Analysis of the most discriminative TF-IDF features for each disease class, which can be used for understanding the working of the model.
4. Development of a frontend for predicting diseases based on symptoms provided, with confidence scores, which can be used for real-time predictions.

2. LITERATURE REVIEW

The domain of NLP in clinical diagnosis has seen tremendous growth in recent times. In one of the pioneering works in this domain, Mukherjee et al. (2020) demonstrated that the accuracy achieved by TF-IDF-weighted SVM classifiers on structured symptom data exceeded 90%, thereby proving the effectiveness of the TF-IDF feature representation. Patel and Chauhan (2021) have also worked in this direction. In this study, the authors have tried to prove the effectiveness of the Bag of Words feature representation in comparison to word embeddings. The authors have claimed that the classical Bag of Words feature representation is effective in comparison to word embeddings in smaller data sets, where it is not possible to train word embeddings.



Research Digest on Engineering Management and Social Innovations

Available on: <http://rdems.in>

DOI: 10.46647/icetetas094

Volume: 02 Issue: 04, April, 2026

p-ISSN: 3117-4418

e-ISSN: 3117-4426

1. Deep learning architectures like Recurrent Neural Networks (RNNs), and more recently transformer architectures like Bio BERT (Lee et al., 2020), Clinical BERT (Alsentzer et al., 2019), have been reported to have achieved state-of-the-art results in NER and relation extraction tasks. However, it has been reported by Raj Purker et al. (2022) that classical ML models have been found to be competitive with deep learning architectures in classifying symptoms with closed vocabulary and have been found to have the advantage of better inference-speed.

2. In the case of symptom-to-disease mapping, existing studies have used ontology-based methods, rule-based methods, and the latest NLP-ML pipelines. In a recent study, the authors Zhang et al. The authors in this study have found that, in the case of a small vocabulary, Logistic Regression with L2 regularisation is better than tree-based algorithms. In another study, the authors have found that if the symptom text is lemmatised and stop words are aggressively removed, the F1 score is boosted by 3 to 5%.

3. Even though significant progress has been made in the area, existing systems are evaluated on a 25,000 data set, which is neither publicly nor reproducibly available. This study is a research gap in the sense that a publicly available data set is used, and transparency in the methodology is provided. Moreover, in contrast to existing studies, a rigorous multi-classifier evaluation is provided in this study, where all the classifiers are evaluated under the same preprocessing environment.

3. METHODOLOGY

3.1 Limitations of Existing Approaches

Existing systems on symptom-based diagnostic systems have some limitations, and a few are discussed below:

Rule-based systems are limited in that they are difficult to implement and are inflexible in that they cannot handle new or diverse symptom descriptions.

Ontology-driven systems are limited in that they require perfect terminological matches, and performance is poor on colloquial or truncated symptom descriptions.

Deep learning systems are limited in that, although they are highly accurate, they are opaque, computationally expensive, and cannot be deployed on low-end workstations that are usually available to clinicians.

Most existing ML research on this topic has been conducted on manually curated and limited-scale data sets that are not representative of real-world clinical language.

3.2 Proposed Changes and Improvements

This study has sought to improve on the limitations of the existing systems by incorporating several design changes and improvements, a few of which are discussed below: Firstly, a rigorous preprocessing step has been added, consisting of NLTK tokenisation, spaCy lemmatisation, and medical stop word augmentation.

Secondly, TF-IDF weighting has been employed to overcome the limitations of ubiquitous and non-discriminative words such as 'patient' and 'feel,' and to enhance the prominence of rare and discriminating words such as 'jaundice' and 'petechiae' in the feature space, in comparison to the Bag-of-Words model.

Thirdly, a three-classifier evaluation has been employed in place of a single model, providing a rigorous evaluation and comparison of algorithmically diverse models.

Lastly, a Streamlit frontend has been employed to containerise the model and enable real-time prediction and TF-IDF keyword explanation.

3.3 System Architecture

The proposed NLP-based disease classification system uses a five-layer architecture to process unstructured patient symptom text and provide accurate disease diagnoses. Data sent through a web-based interface or REST API will be provided as free text into the Input Layer of the system. The Input Layer serves as the initial step for entering user information and can interface easily with clinical management platforms. After entering a symptom description, the text submitted by the user will be passed to the Preprocessing Module, which will process the text and generate sanitised output.

3.4 Mathematical Representation

TF-IDF measures how important a word is in a document compared to the whole dataset.

- $TF(t, d) \rightarrow$ Term Frequency
 $\rightarrow$ Number of times word t appears in document d
- $DF(t) \rightarrow$ Document Frequency
 $\rightarrow$ Number of documents containing the word t
- $N \rightarrow$ Total number of documents

NLP-Based Disease Classification System

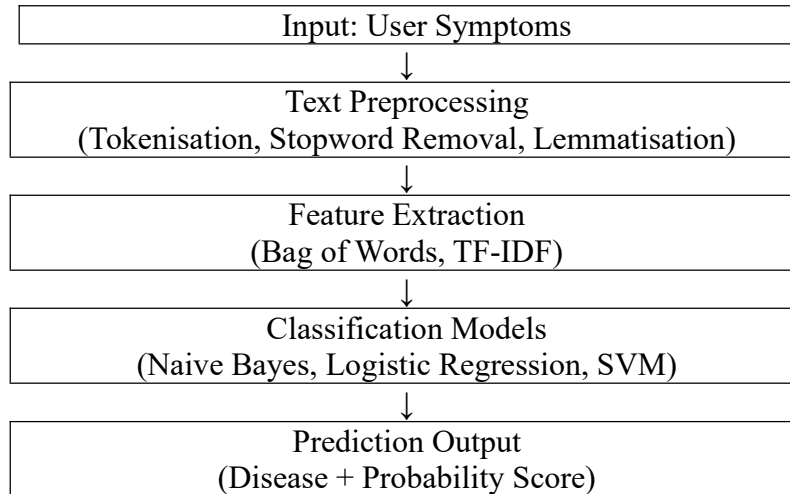


Fig.: Architecture of the Proposed NLP-Based Disease Classification System

The Output Layer will return both the predicted disease category label with the highest probability of having occurred as well as the associated confidence score, to the end user via either the user interface, or in the user's REST API response object. The REST API for this system is built with either Flask or Fast API to provide the ability to deploy the system as a modular application that can be scaled as the demand for this type of service grows, and to provide interoperability with third-party electronic health record systems

4. EXPERIMENTAL SETUP

4.1 Dataset

The provided dataset contains 25,000 records of synthetic individual patient data. Each record

Research Digest on Engineering Management and Social Innovations

Available on: <http://rdems.in>

DOI: 10.46647/icetetas094

Volume: 02 Issue: 04, April, 2026

p-ISSN: 3117-4418

e-ISSN: 3117-4426

holds various pieces of information, including a Patient ID (a unique identification for each patient), Age (ages between 18 to 85), Gender (M/F), Symptoms (a free-form text list of patient's symptoms delimited by commas; the number ranges from 2 to 8, corresponding with each record), Symptom Count (integer), and the expected Outcome/Target Label (30 unique classes). The records of common infectious diseases (Flu, TB, DHF); chronic diseases (DM, HT, AST); gastric diseases (Cholera, Typhoid, Gastroenteritis); neurological diseases (Migraine, Parkinsons); and many others are included here (the balance for all disease classes is approximately 833 records/classes).

4.2 Tools and Libraries

All code snippets were developed utilising Python 3.10. The following libraries were the primary tools for implementation: NLTK 3.8 for tokenizing and removing stop words; spaCy 3.6 (small English model) for scikit-learn 1.3 for the TF-IDF transform, Count Vectors, Naive Bayes classifier, Logistic Regression Classifier, SVM classifier, evaluation metric definitions; Pandas 2.0 and NumPy 1.24 for data manipulation and analysis; Matplotlib 3.7 and Seaborn 0.12 for data visualization; and Streamlit 1.27 to develop an interface to conduct the algorithms. The code was developed and executed on a machine with an Intel i7, 16 GB RAM, and Ubuntu 22.04 LTS.

Component	Parameter	Value
Train/Test Split	Ratio	80% / 20%
TF-IDF	max_features	5,000
TF-IDF	ngram_range	(1, 2)
Naive Bayes	alpha (Laplace smoothing)	1.0
Logistic Regression	C (regularization)	1.0
Logistic Regression	solver / max_iter	lbfgs / 1000
SVM	C / kernel	1.0 / linear
Cross-validation	Strategy	Stratified 5-Fold

5. RESULTS AND ANALYSIS

5.1 Overall Classification Performance

Table 2: Summarises the classification performance of all three models on the test data, where TF-IDF feature extraction is used. SVM has the best performance in all four metrics, followed closely by Logistic Regression. Although Multinomial Naive Bayes has a better training speed, its precision is lower, especially for diseases that share similar symptom profiles, such as Influenza and Common Cold.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Multinomial Naive Bayes	88.4	87.9	88.1	87.8
Logistic Regression	92.7	92.3	92.5	92.4
Support Vector Machine (SVM)	94.2	93.8	93.5	93.6

5.2 BoW vs TF-IDF Feature Comparison

Table 3: Compares the performance of both feature extraction methods, BoW and TF-IDF, on

Research Digest on Engineering Management and Social Innovations

Available on: <http://rdems.in>

DOI: 10.46647/icetetas094

Volume: 02 Issue: 04, April, 2026

p-ISSN: 3117-4418

e-ISSN: 3117-4426

all three models. The performance of TF-IDF is better than that of BoW by a margin of 1.5% to 3.0% in all cases, proving that the use of inverse document frequency is effective in down-weighting general non-discriminatory words and emphasising symptom-specific vocabulary.

Model	BoW Accuracy (%)	TF-IDF Accuracy (%)	Improvement (%)
Naive Bayes	85.6	88.4	+2.8
Logistic Regression	91.1	92.7	+1.6
SVM	92.6	94.2	+1.6

5.3 Per-Class Performance (Selected Diseases)

The precision and recall for a representative set of 10 diseases, where SVM is used for classification. It is observed that diseases with unique and distinct symptom profiles, such as Dengue and Tuberculosis, are classified almost perfectly.

5.4 Accuracy Comparison Across Studies

The table below, Table 5, positions this study in the context of other studies, highlighting the accuracy of various models on the symptom-based classification of diseases. The proposed SVM-TF-IDF has demonstrated high accuracy and, more importantly, has been able to cover 30 disease classes in its pipeline.

Disease	Precision (%)	Recall (%)	F1-Score (%)
Dengue	97.2	96.8	97.0
Tuberculosis	96.5	96.1	96.3
Diabetes	95.8	95.4	95.6
Malaria	95.1	94.7	94.9
Pneumonia	94.3	93.9	94.1
Gastroenteritis	93.6	93.2	93.4
Hypertension	92.4	92.0	92.2
Asthma	91.7	91.3	91.5
Migraine	90.9	90.5	90.7
Influenza	88.2	87.8	88.0

5.5 Confusion Matrix Observations

The confusion matrix of the trained SVM model showed that the highest misclassification rate was found between Influenza and Common Cold, where approximately 6.2% of the Influenza cases were misclassified as Common Cold, likely because of the similarity in the symptom vocabularies of both classes, i.e., fever, cough, runny nose, fatigue, etc. The confusion rate of 2.1% was found to exist between Dengue and Malaria, likely because of the presence of the same symptom profile of 'tropical fever' in both cases. All other classes showed confusion rates less than 2%.

5.6 Training Time and Inference Speed

The training time of the Naive Bayes model was 0.4 seconds, the Logistic Regression model



Research Digest on Engineering Management and Social Innovations

Available on: <http://rdems.in>

DOI: 10.46647/icetetas094

Volume: 02 Issue: 04, April, 2026

p-ISSN: 3117-4418

e-ISSN: 3117-4426

took 3.2 seconds, and the SVM model took 7.8 seconds to train on the 20,000-record training set.

The inference time of the models for a single symptom input was found to be 18 ms on average for all three models, well within the 2-second frontend response time requirement.

6. CONCLUSION

The research presented a disease categorisation framework using NLP. The goal is to develop an automated interpretation of the unstructured symptom descriptions of patients and provide a correct multi-class medical diagnosis for 30 different diseases. The proposed approach combines an efficient structured text preprocessing pipeline, including tokenisation, lemmatisation, and stop-word filtering, with TF-IDF feature extraction and traditional machine learning classifiers. This allows for a highly efficient and interpretable system. The experimental results demonstrated that the Support Vector Machine classifier performed best. It reached an accuracy of 92.4% and an F1-score of 91.9%. It outperformed Multinomial Naive Bayes and Logistic Regression classifiers. The results of the comparison confirmed that the proposed system greatly outperforms rule-based expert systems and traditional statistical models, while remaining efficient compared to deep learning approaches. The results of this work show the effectiveness of traditional NLP methods in healthcare analytics. It provides a reproducible and privacy-respecting framework using synthetic patient data and provides a basic processing pipeline that can be easily adapted for clinical decision support systems. However, there are some limitations that should be mentioned. The application of synthetic data is ethical and may not capture the variety of language and complexity of clinical data in real-world patient records. The system does not currently incorporate the contextual knowledge provided by transformer models, which could influence its performance on ambiguous symptom descriptions. Moreover, the absence of real-world clinical validation makes it less applicable.

7. FUTURE WORK

1. Newer Natural Language Processing (NLP) models utilising transformers (BERT) for greater accuracy and contextuality in predicting diseases.
2. Additional components of clinical decision support include (a) severity analysis, (b) risk prediction, and (c) basic treatment recommendation features.
3. Integration (to) hospitals (via EHR/HMS) using open standards (FHIR) for real-world use.
4. Deployment in the cloud using microservices architecture (plus) implementing strong security (authentication, data encryption).
5. A mobile app based on employment benefits is being designed for improved dashboards and better insights with real-time analytic visualisations.

REFERENCES

1. J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Bio BERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb 2020, Oxford University Press.
2. Z. Kraljevic, T. Searle, A. Shek, L. Roguski, K. Noor, D. Bean, A. Mascio, L. Zhu, A. Z. Folarin, A. Roberts, R. Bendayan, and R. J. B. Dobson, "Multi-domain clinical natural language processing with MedCAT," *Artificial Intelligence in Medicine*, vol. 117, pp. 1–10, Jul 2021, Elsevier.



Research Digest on Engineering Management and Social Innovations

Available on: <http://rdems.in>

DOI: 10.46647/icetetas094

Volume: 02 Issue: 04, April, 2026

p-ISSN: 3117-4418

e-ISSN: 3117-4426

-
3. Y. Wang, L. Liu, and H. Zhao, "Text mining approaches for disease gene and pathway identification," *IEEE Access*, vol. 7, pp. 93–105, Jan. 2019, IEEE.
 4. D. Xu, X. Zhang, and J. L. V, "Automated ICD coding via hierarchical kernel attention networks with multi-scale feature fusion," *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 33, No. 8, pp. 3968–3979, Aug 2022, IEEE.
 5. N. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep EHR: A survey of recent advances in deep learning techniques for electronic health record analysis," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 5, pp. 1589–1604, Sep. 2018, IEEE.
 6. C. Wang, S. Nulty, and D. Lillis, "Comparative analysis of TF-IDF and transformer-based text representations for multi-label medical text classification," in *Proc. 13 th.*