International Journal of Artificial Intelligence in Medical Issues

Volume 3 Issue 2
ISSN 3025-4167
https://doi.org/10.56705/ijaimi.v3i2.361

Research Article

Comparative Study of Machine Learning Methods for Disease Classification Based on Natural Language Symptom Descriptions

Ery Setiyawan Jullev Atmadji [1,*]; Adityo Permana Wibowo [2]; Edi Faizal [3]
[1] Politeknik Negeri Jember, Kabupaten Jember, Jawa Timur 68121, Indonesia, ery@polije.ac.id
[2] Universitas Teknologi Yogyakarta, Daerah Istimewa Yogyakarta 55285, Indonesia, adityopw@uty.ac.id
[3] Universitas Teknologi Digital Indonesia, Daerah Istimewa Yogyakarta 55198, Indonesia, edifaizal@utdi.ac.id

Correspondence should be addressed to Ery Setiyawan Jullev Atmadji; ery@polije.ac.id

Received 04 September 2025; Revised 10 September 2024; Accepted 25 October 2024; Published 30 November 2025

Abstract:
The growing demand for remote healthcare solutions has increased the importance of efficient disease diagnosis based on textual symptom descriptions. This study explores the application of machine learning models Multinomial Naive Bayes, Random Forest, and Support Vector Machine (SVM) to classify 24 different diseases from natural language symptom inputs. Utilizing a dataset of 1,200 balanced samples and TF-IDF for feature extraction, we trained and evaluated the models using both accuracy and cross-validation metrics. Among the models, SVM achieved the highest test accuracy of 97.5% and demonstrated consistent performance across all disease categories. These findings underscore the potential of classical machine learning approaches in enhancing digital diagnostic tools, particularly for early screening in telemedicine applications. Future work could extend this study by integrating deep learning architectures and multilingual capabilities to accommodate broader and more diverse healthcare scenarios.

Keywords: Natural language processing, Disease classification, Symptom description, Machine learning, Support Vector Machine, Naive Bayes, Random Forest, TF-IDF, Text classification, Telemedicine.

Dataset link: https://www.kaggle.com/datasets/niyarrbarman/symptom2disease/versions/1

1. Introduction

The accurate and timely classification of diseases based on patients' symptom descriptions is a pressing challenge and an active research area in the domain of digital health. With the proliferation of natural language inputs via telemedicine platforms, chatbots, and digital assistants, the necessity to transform unstructured symptom narratives into meaningful diagnostic predictions has become increasingly important. In this context, Natural Language Processing (NLP) combined with machine learning (ML) models offers a promising solution to support early and remote diagnosis.

Despite rapid advancements in medical NLP applications, a research gap persists in addressing multi-class disease classification based solely on short and ambiguous textual symptom descriptions. Prior works often focus on binary or limited disease sets, underrepresenting the complexity of real-world symptoms and their linguistic variability [1], [2], [3]. Furthermore, while pre-trained language models and deep learning architectures dominate recent studies [4],

[5], traditional machine learning methods such as Naive Bayes, Support Vector Machines (SVM), and Random Forest remain competitive and interpretable alternatives, particularly in resource-constrained settings [6], [7].

Current state-of-the-art research highlights the application of interpretable ML techniques to various medical text datasets for clinical support, demonstrating potential in early disease screening, triage systems, and health monitoring tools [8]. However, the performance of classical ML models for multi-class disease classification using short symptom texts remains underexplored.

This study aims to investigate the effectiveness of three widely adopted ML classifiers Naive Bayes, Random Forest, and SVM in classifying 24 distinct diseases based on 1,200 short natural language symptom descriptions. The focus lies in evaluating the accuracy, generalizability, and robustness of these models in distinguishing diseases with overlapping symptoms.

From an empirical standpoint, challenges include limited context within user symptom descriptions, semantic overlap between diseases, and lexical sparsity. This study addresses these issues through pre-processing, TF-IDF vectorization, and stratified model evaluation, contributing both practical insights and reproducible methodology for future research and implementation.

2. Method

Research Design:

This study adopts a quantitative research design grounded in supervised machine learning techniques to compare three classifiers Naive Bayes, Support Vector Machine (SVM), and Random Forest for multiclass disease classification based on free-text symptom descriptions. The workflow follows a structured pipeline: data loading and exploration → text pre-processing → feature extraction → model training & evaluation [9], [10], [11]. This design allows for reproducible comparisons among models under identical data and feature settings.

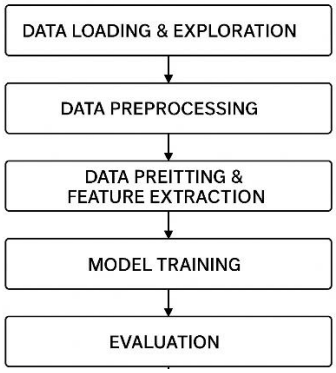

Figure 1: Research Workflow

Data Loading & Exploration:

We began by loading the dataset of 1,200 samples and two columns (`label`, `text`). Basic statistics such as number of classes (24 diseases) and distribution of records were obtained. Descriptive metrics on text length

(characters) and word count were also computed to understand text variability. This preliminary step aligns with practices in text classification pipeline research that emphasize early exploration of input text distributions [12].

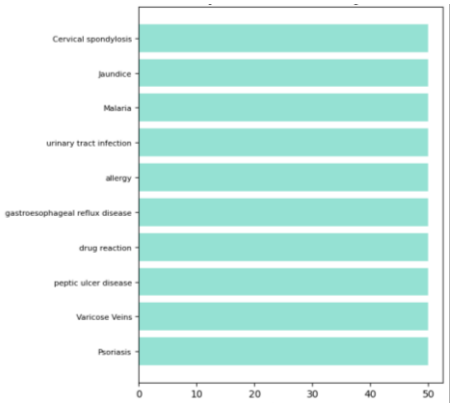

Figure 2: Top 10 Disease Distribution.

Pre-processing Text:

To prepare the free-text symptom descriptions for modelling, the following steps were applied:

- Lowercasing all characters.
- Removing non-alphabetic characters via regex.
- Trimming extra whitespace.
- Encoding labels using LabelEncoder.

Cleaning the text in such a way reduces noise and enhances consistency for feature extraction. This approach echoes recommended pre-processing stages in recent machine-learning pipelines for text classification [13], [14], [15].

Feature Extraction:

For turning cleaned text into numeric features usable by ML models, we applied the TF-IDF vectorizer (with `max_features=1000` and `ngram_range=(1,2)`) [10], [16], [17]. The transformation is represented mathematically as:

$$tfidf_{i,j} = \frac{tf_{i,j}}{df_j} \times log\frac{N}{1+df_j} \quad (2)$$

Where $tf_{i,j}$ is term-frequency of term j in document i, df_j is document-frequency of term j and $N$is total number of documents. Choosing TF-IDF rather than simple bag-of-words improves weighting of rare yet informative terms; this decision is supported by comparative studies in text classification literature [18], [9], [19].

Splitting Data:

The dataset was stratified into training (80 %) and testing (20 %) sets to preserve class distribution across disease labels. Stratification is important in multiclass settings to avoid sampling bias and to ensure generalizability of results [15].

Model Development and Evaluation

Three classifiers were trained with identical feature sets and splits:

- Naive Bayes (MultinomialNB with α=0.1)
- Random Forest (100 trees, random_state=42)
- Support Vector Machine (kernel='linear', C=1.0)

This comparative setup enables direct evaluation of how each algorithm handles the same text-based classification task [20], [21], [22].

Evaluation:

Models were evaluated on the test set using accuracy, precision, recall, and F1-score standard metrics in classification tasks [23], [24]. Additionally, 5-fold cross-validation on training data was used to estimate model stability.

$$Mean\ CV\ Accuracy = \frac{1}{k}\sum_{i=1}^{k} Acc_i \quad (3)$$

where $k = 5$. The selection of these metrics and evaluation protocol aligns with best practices in text classification and ML research [14].

3. Result and Discussion

The evaluation of machine learning models for disease classification based on natural language symptom descriptions yielded highly promising results. The dataset comprised 1,200 symptom descriptions distributed evenly across 24 disease classes. Each class was well-represented, allowing for a balanced analysis and reliable comparison of models.

The three models tested were Multinomial Naive Bayes, Random Forest, and Support Vector Machine (SVM), each trained on TF-IDF-transformed symptom text data. The performance evaluation was conducted on a hold-out test set comprising 20% of the data, with additional validation via 5-fold cross-validation.

The SVM model outperformed the others with a test accuracy of 97.5%, followed by Naive Bayes at 95.0%, and Random Forest at 93.75% (**Figure 3**). These results are further corroborated by cross-validation scores, with SVM achieving a mean CV score of 95.94%, again the highest among the three models. This superior performance can be attributed to the SVM's capability to handle high-dimensional spaces and its effectiveness in separating classes with clear decision boundaries in sparse data representations like TF-IDF.:

Table 1. Training and Validation Performance per Epoch

Model	Accuracy	Mean CV Score	Std. Dev. CV
Naive Bayes	0.95	0.9469	0.0212
Random Forest	0.9375	0.9323	0.0329
SVM	0.975	0.9594	0.0283

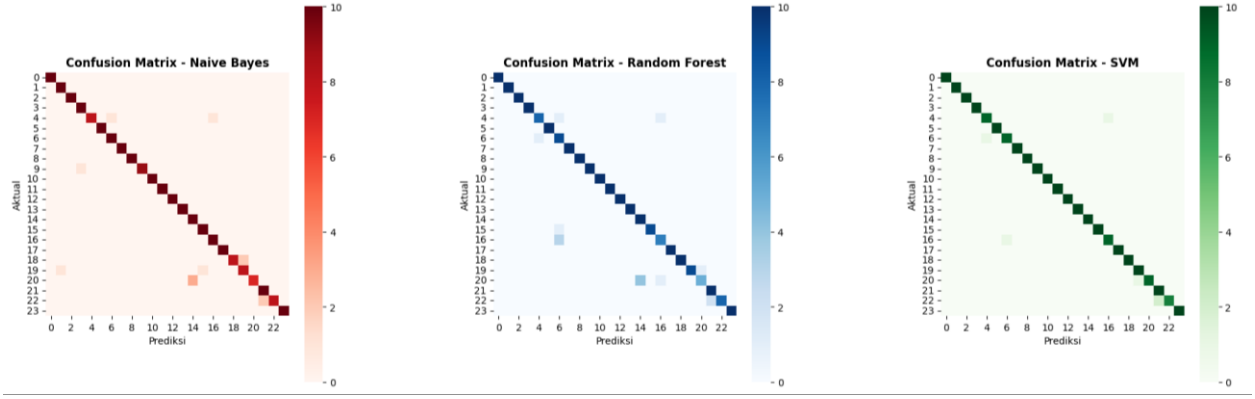

Figure 3. Examination of the Confusion Matrices.

A closer examination of the confusion matrices (**Figure 3**) reveals that the SVM model maintains a consistently high true positive rate across almost all disease categories, with very few misclassifications. In contrast, Random Forest and Naive Bayes showed slight confusion between classes with overlapping symptoms, such as dengue and malaria or typhoid and peptic ulcer disease.

The classification reports further validates these observations. SVM achieved perfect or near-perfect precision and recall across most classes, whereas Random Forest, despite strong general performance, displayed lower recall in a few disease classes (e.g., drug reaction, dengue). Naive Bayes, although slightly behind SVM in overall accuracy, demonstrated robust generalization due to its probabilistic nature, which suits text classification tasks well.

The distribution of diseases in the dataset was uniformly balanced (**Figure 2**), which eliminates bias in training and allows performance metrics to truly reflect model capabilities without data imbalance artifacts. This reinforces the reliability of the performance outcomes.

In conclusion, this experiment demonstrates the effectiveness of using TF-IDF features combined with SVM for classifying diseases from textual symptom descriptions. Given the growing importance of remote diagnostics and the need for scalable, language-based health applications, the results support integrating such models into telemedicine platforms for preliminary diagnostic assistance.

Future research may include evaluating deep learning-based models such as BERT or LSTM for even greater semantic understanding, handling multilingual symptom inputs, and assessing real-world deployment robustness.

4. Conclusion

This study presents a comprehensive evaluation of machine learning methods Multinomial Naive Bayes, Random Forest, and Support Vector Machine for classifying diseases from natural language symptom descriptions. Utilizing TF-IDF vectorization and a balanced dataset covering 24 distinct diseases, all models demonstrated high classification accuracy, with SVM achieving the best performance at 97.5% test accuracy and 95.94% average cross-validation accuracy.

The findings confirm the viability of using classical machine learning approaches for text-based medical classification tasks. The exceptional performance of SVM suggests its potential as a core component in digital diagnostic tools, particularly in telemedicine and early disease screening applications.

By converting unstructured symptom narratives into structured predictions, this research contributes to advancing automated health support systems. Future studies should aim to scale this approach with larger, real-world datasets, incorporate deep learning models, and consider contextual factors such as temporal symptom progression and patient history to further enhance prediction accuracy and applicability.

References:

[1] A. Jerfy, O. Selden, and R. Balkrishnan, "The Growing Impact of Natural Language Processing in Healthcare and Public Health," *Inq. J. Heal. Care Organ. Provision, Financ.*, vol. 61, Jan. 2024, doi: 10.1177/00469580241290095.

[2] S. M. A. Rahman, S. Ibtisum, E. Bazgir, and T. Barai, "The Significance of Machine Learning in Clinical Disease Diagnosis: A Review," *Int. J. Comput. Appl.*, vol. 185, no. 36, pp. 10–17, 2023, doi: 10.48550/arXiv.2310.16978.

[3] M. M. Ahsan, S. A. Luna, and Z. Siddique, "Machine-Learning-Based Disease Diagnosis: A Comprehensive Review," *Healthcare*, vol. 10, no. 3, p. 541, Mar. 2022, doi: 10.3390/healthcare10030541.

[4] F. Sogandi, "Identifying diseases symptoms and general rules using supervised and unsupervised machine learning," *Sci. Rep.*, vol. 14, no. 1, p. 17956, Aug. 2024, doi: 10.1038/s41598-024-69029-8.

[5] T. Ling, L. Jake, J. Adams, K. Osinski, X. Liu, and D. Friedland, "Interpretable machine learning text classification for clinical computed tomography reports – a case study of temporal bone fracture," *Comput. Methods Programs Biomed. Updat.*, vol. 3, p. 100104, 2023, doi: 10.1016/j.cmpbup.2023.100104.

[6] A. Fuster-Palà, F. Luna-Perejón, L. Miró-Amarante, and M. Domínguez-Morales, "Optimized Machine Learning Classifiers for Symptom-Based Disease Screening," *Computers*, vol. 13, no. 9, p. 233, Sep. 2024, doi: 10.3390/computers13090233.

[7] Y. Wang, J. Zhong, and R. Kumar, "A Systematic Review of Machine Learning Applications in Infectious Disease Prediction, Diagnosis, and Outbreak Forecasting." Apr. 15, 2025, doi: 10.20944/preprints202504.1250.v1.

[8] N. Ghaffar Nia, E. Kaplanoglu, and A. Nasab, "Evaluation of artificial intelligence techniques in disease diagnosis and prediction," *Discov. Artif. Intell.*, vol. 3, no. 1, p. 5, Jan. 2023, doi: 10.1007/s44163-023-00049-5.

[9] A. Ranjan, "An Ensemble Tf-Idf Based Approach to Protein Function Prediction via Sequence Segmentation," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 19, no. 5, pp. 2685–2696, 2022, doi: 10.1109/TCBB.2021.3093060.

[10] S. M. M. Hossain, "TF-IDF feature-based spam filtering of mobile SMS using a machine learning approach," *Applied Intelligence for Industry 4.0*. pp. 162–175, 2023.

[11] J. W. Sun, "Text Classification Algorithm Based on TF-IDF and BERT," *Proceedings - 2022 11th International Conference of Information and Communication Technology, ICTech 2022*. pp. 533–536, 2022, doi: 10.1109/ICTech55460.2022.00112.

[12] H. Allam, L. Makubvure, B. Gyamfi, K. N. Graham, and K. Akinwolere, "Text Classification: How Machine Learning Is Revolutionizing Text Categorization," *Information*, vol. 16, no. 2, p. 130, Feb. 2025, doi: 10.3390/info16020130.

[13] C. Rodríguez-Penagos *et al.*, "FBM: Combining lexicon-based ML and heuristics for Social Media Polarities," 2013.

[14] A. Ise, "Machine Learning Pipeline for multi-class text Classification," *Int. J. Eng. Appl. Sci. Technol.*, vol. 7, no. 2, pp. 64–69, 2022.

[15] M. Siino, I. Tinnirello, M. La Cascia, and B. -Delft, "The Text Classification Pipeline: Starting Shallow, going Deeper From Foundations to GPT in Text Classification: A Comprehensive Survey on Current Approaches and Future Trends," 2024, doi: 10.1561/XXXXXXXXX.Marco.

[16] G. Popoola, "Sentiment Analysis of Financial News Data using TF-IDF and Machine Learning Algorithms," *2024 IEEE 3rd International Conference on AI in Cybersecurity, ICAIC 2024*. 2024, doi: 10.1109/ICAIC60265.2024.10433843.

[17] S. M. M. Hossain, K. M. A. Kamal, A. Sen, and I. H. Sarker, *TF-IDF Feature-Based Spam Filtering of Mobile SMS Using a Machine Learning Approach*. 2023.

[18] A. Occhipinti, L. Rogers, and C. Angione, "A pipeline and comparative study of 12 machine learning models for text classification," *Expert Syst. Appl.*, vol. 201, p. 117193, Sep. 2022, doi: 10.1016/j.eswa.2022.117193.

[19] K. Yusupov, "Comparative Analysis of Machine Learning and Deep Learning Models for Email Spam Classification Using TF-IDF and Word Embedding Techniques," *Lecture Notes on Data Engineering and Communications Technologies*, vol. 231. pp. 114–122, 2025, doi: 10.1007/978-3-031-76452-3_11.

[20] K. S. Gill, "Hypothesis Testing of Gaussian Naïve Bayes Classifier for Liver Disease Classification," *2023 2nd International Conference on Futuristic Technologies, INCOFT 2023*. 2023, doi: 10.1109/INCOFT60753.2023.10425015.

[21] C. R. Dhivyaa, "Skin lesion classification using decision trees and random forest algorithms," *J. Ambient Intell. Humaniz. Comput.*, 2020, doi: 10.1007/s12652-020-02675-8.

[22] A. S. Khan, "Integrating BERT Embeddings with SVM for Prostate Cancer Prediction," *Proceedings - 6th International Conference on Electrical Engineering and Information and Communication Technology,*

ICEEICT 2024. pp. 574–579, 2024, doi: 10.1109/ICEEICT62016.2024.10534547.

[23] H. Azis, M. Abdullah, S. Ismail, and ..., "A Comparative Study of YOLO Models for Enhanced Vehicle Detection in Complex Aerial Scenarios," *2025 19th Int. …*, 2025.

[24] A. R. Manga, H. Azis, F. Fattah, Y. Salim, and ..., "ResNet-50 for Flower Image Classification: A Comparative Study of Segmentation and Non-Segmentation Approaches," *2025 19th …*, 2025.