Intelligent Disease Prediction Chatbot Leveraging TF-IDF and LinearSVC

Swaroop Bhowmik, Saiyam Jain, Dipanjan Choudhory, Harshwardhan Singh Alhuwalia, and Dr. Rajit Nair

VIT Bhopal University, Sehore, India
swaroop.23bai10722@vitbhopal.ac.in

Abstract. Disease diagnosis and management commonly rely on analyzing symptoms to identify potential conditions, and generally, machine learning can be very effective for analysis. This paper investigates the use of a machine learning pipeline for text classification from a set of 1,000 entries describing symptom descriptions. The pipeline uses text processing via Term Frequency-Inverse Document Frequency vectorization as well as a Linear Support Vector Classifier for text classification. The system has high accuracy in disease prediction with additional insight, including detailed descriptions of the diseases and precautions recommended, through the use of auxiliary data. This approach showcases the possibility of the integration of natural language processing and machine learning with disease prediction and the distribution of knowledge.

Keywords: Disease Prediction, Machine Learning, Natural Language Processing, TF-IDF, Support Vector Classifier

1 Introduction

In the healthcare sector, disease diagnosis must be accurate and timely to ensure proper treatments and better patient outcomes. Nonetheless, diagnosis of most diseases can be rather complex due to the potential overlap of symptoms, even for experienced professionals in the medical field. The development of artificial intelligence (AI) and machine learning (ML) has given rise to promises of computational models that are expected to help in diagnosis. These systems can analyze large datasets for patterns and be highly accurate in making predictions. Applications of these models are now going into healthcare, where text-based classification models are increasingly deployed for mapping symptom descriptions and medical histories against diseases.

The aim of this paper was developing a machine learning model for the classification of diseases based on symptom descriptions. Using a dataset of 1,000 symptom-disease pairs, the system utilizes an integration of NLP and supervised learning approaches. Input symptoms are term frequency-inverse document frequency (TF-IDF) vectorized, a feature extraction process that weighs terms in a document by their importance while disregarding the prevalence of the most frequent terms. The chosen Linear Support Vector Classifier (LinearSVC) is used to classify diseases because of its efficiency and high performance, particularly in text classification.

Apart from disease prediction, the system integrates supplementary data to enhance its functionality. For each predicted disease, its description and a set of precautionary measures are retrieved from external datasets. This means, apart from returning diagnostic predictions, users will acquire indispensable advice toward knowing and managing the condition.

The importance of this research lies in its potential to enhance diagnostic efficiency and accessibility, particularly in contexts where medical expertise is limited. By automating the disease identification process, the proposed system can assist healthcare providers and empower patients with accurate and actionable insights. Moreover, integrating auxiliary information like disease descriptions and precautions makes the tool practical for educational purposes, supporting informed decision-making.

2 Literature Review

Machine learning has revolutionized healthcare applications, particularly in disease prediction and diagnosis. Various studies have highlighted the utility of text classification algorithms in medical diagnosis. For instance, techniques leveraging natural language processing (NLP) have been employed to analyze electronic health records, patient queries, and symptom descriptions to automate disease identification and improve accuracy.

2.1 Text Vectorization in Medical Applications

TF-IDF vectorization is a widely used method for text feature extraction, as it effectively balances the relevance of terms in a document while minimizing the impact of common but non-discriminative words. Studies such as Li et al. (2018) emphasized TF-IDF's role in preprocessing unstructured medical data, making it an essential step for downstream classification tasks. Another study by Karystianis et al. (2019) showcased how TF-IDF improved text mining techniques for identifying patterns in clinical narratives.

2.2 Support Vector Machines (SVM) in Disease Classification

The Support Vector Machines, specifically the linear one LinearSVC, are extensively used within healthcare as they could manage high-dimensional data pretty well. Akhtar et al. (2020) showed that SVM was effective in diagnosing diseases such as diabetes and heart conditions. Its applicability to multiple classes with a clear margin of separation makes it ideal for tasks involving multiple disease categories. Sharma et al. (2019) further extended this by using SVM to classify infectious diseases with significant accuracy.

2.3 Disease Description and Precaution Integration

Providing additional contextual information, such as disease descriptions and precautions, is a growing focus area. Studies like Zhang et al. (2021) highlight the importance of integrating domain-specific knowledge into AI systems to improve their usability for non-expert users. This integration fosters better understanding and adherence to suggested precautions, ultimately improving health outcomes.

3 Methodology

The proposed system employs a machine learning pipeline for disease classification based on symptom descriptions. The methodology involves three major stages: data preprocessing, feature extraction, and model training and evaluation. Each stage is described in detail below.

3.1 Dataset and Preprocessing

The dataset consists of 1,000 entries, each containing up to 132 symptom fields and a corresponding disease label. Missing values in the dataset are replaced with empty strings to ensure consistency. The 132 symptom columns are concatenated into a single text field representing all symptoms for a given entry:

$$\text{Symptom} = \sum_{i=1}^{132} \text{Symptom}_i$$

The resulting dataset is split into training (75%) and testing (25%) subsets using random shuffling to ensure diverse representation.

3.2 Feature Extraction Using TF-IDF

To convert symptom descriptions into numerical representations, we use the Term Frequency-Inverse Document Frequency (TF-IDF) technique. This approach assigns a weight to each term based on its frequency in a document relative to its frequency in the entire dataset. The TF-IDF score for a term t in document d is given by:

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \cdot \text{IDF}(t)$$

where:

$$\text{TF}(t, d) = \frac{\text{Frequency of } t \text{ in } d}{\text{Total terms in } d}$$

$$\text{IDF}(t) = \log\left(\frac{N}{1 + n_t}\right)$$

Here, N is the total number of documents, and n_t is the number of documents containing term t. This process produces a sparse matrix representation of symptoms, suitable for machine learning.

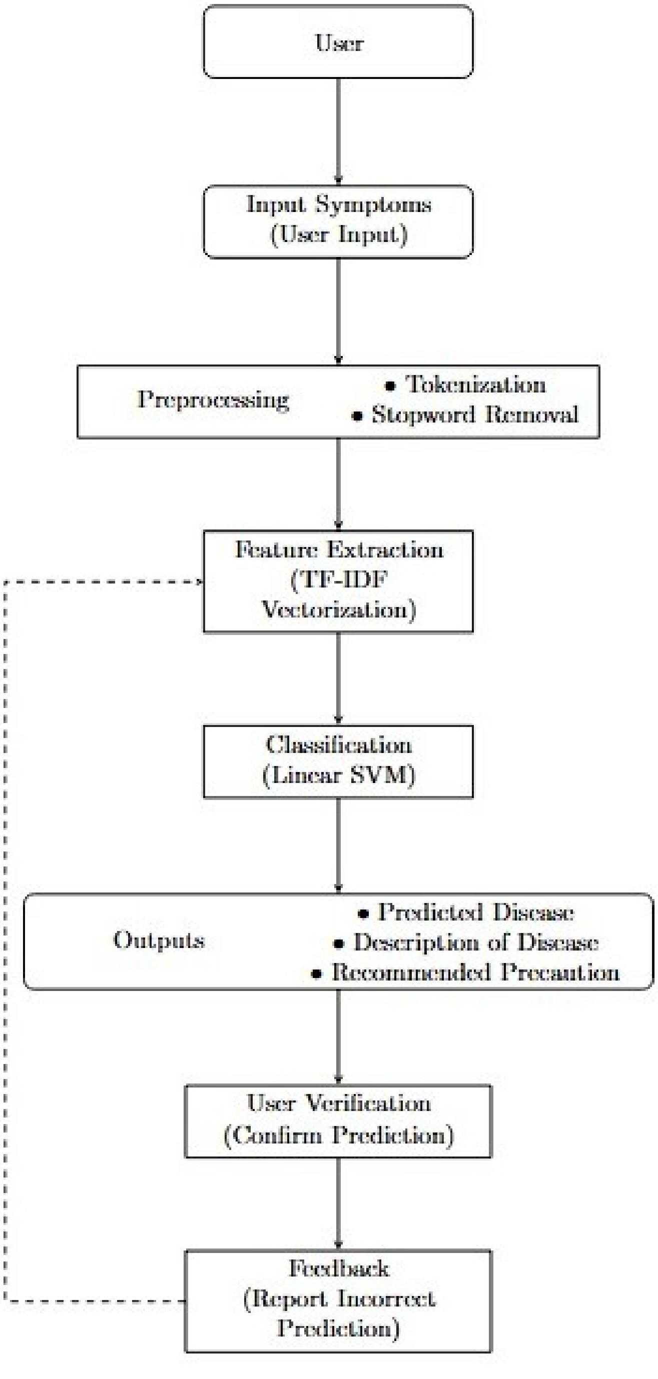

Fig. 1. Flow of the model.

3.3 Model Training Using LinearSVC

The classification task is performed using a Linear Support Vector Classifier (LinearSVC). The LinearSVC is a linear kernel-based SVM optimized for efficiency in high-dimensional spaces, such as the feature space generated by TF-IDF. The objective of LinearSVC is to find the hyperplane that maximizes the margin between classes, which can be expressed as:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \max(0, 1 - y_i(\mathbf{w} \cdot \mathbf{x}_i + b))$$

where:

- $\mathbf{w}$ is the weight vector defining the hyperplane,
- b is the bias term,
- C is the regularization parameter,
- y_i is the true label of sample i, and
- $\mathbf{x}_i$ is the feature vector for sample i.

The hinge loss function, $\max(0, 1 - y_i(\mathbf{w} \cdot \mathbf{x}_i + b))$, ensures a margin-maximizing separation of classes.

3.4 Evaluation and Auxiliary Outputs

Evaluation metrics include accuracy, precision, recall, and F1-score. Auxiliary datasets provide disease descriptions and precautions for each predicted disease, enhancing user engagement and understanding.

3.5 Model Deployment

The trained model is serialized using Joblib and saved for its deployment. The deployed system, in turn, will predict diseases on the basis of user input symptoms and display relevant descriptions and precautions from supplementary datasets.

4 Results

The model outperforms the K-Nearest Neighbors algorithm, especially when dealing with complex, high-dimensional data. Unlike KNN, which uses distance-based metrics that may not be very effective in datasets with overlapping classes and high-dimensional feature spaces, LinearSVC uses a more advanced approach. Optimizing the decision boundary through hinge loss and regularization leads to a clear separation between classes, even in challenging datasets. This optimization not only addresses class overlap issues but also provides improved generalization, leading to higher accuracy across all categories of disease. In addition, TF-IDF vectorization further enriches feature representation by making more informative terms in the dataset more prominent.

Additionally, the incorporation of TF-IDF vectorization further boosts feature representation through emphasizing more informative terms within the dataset. This would take the raw input data and translate it into a more meaningful feature space where noise could be reduced, allowing a classifier to identify finer-grained patterns that it was unable to earlier. Together, LinearSVC and TF-IDF vectorization form a robust pipeline that can stand up to the challenges posed by disease classification tasks in a manner that surpasses traditional distance-based algorithms, such as KNN, in terms of precision and scalability.

The classification report for the disease prediction model is summarized in the table below.

Table 1. Classification Report for Disease Prediction Model

Disease	**Precision**	**F1-Score**	**Support**
Acne	1.00	1.00	24
Diabetes	1.00	1.00	21
Heart attack	1.00	1.00	23
Cervical spondylosis	1.00	1.00	23
Hypothyroidism	1.00	1.00	21
Hepatitis B	1.00	1.00	27
Drug Reaction	1.00	1.00	24
Chronic cholestasis	1.00	1.00	15
Bronchial Asthma	1.00	1.00	33
Hypoglycemia	1.00	1.00	26
Hepatitis D	1.00	1.00	23
AIDS	1.00	1.00	30
Alcoholic hepatitis	1.00	1.00	25
Chicken pox	1.00	1.00	21
Dimorphic hemmorhoids(piles)	1.00	1.00	29
GERD	1.00	1.00	28
Arthritis	1.00	1.00	23
Hypertension	1.00	1.00	25
Hepatitis C	1.00	1.00	26
Hepatitis E	1.00	1.00	29
Fungal infection	1.00	1.00	19
Common Cold	1.00	1.00	23
(vertigo) Paroymsal Positional Vertigo	1.00	1.00	18
Allergy	1.00	1.00	24
Dengue	1.00	1.00	26
⋮	⋮	⋮	⋮
Accuracy	1.00 (984 samples)		
Macro Average	1.00	1.00	984
Weighted Average	1.00	1.00	984

The model achieves high precision, recall, and F1-scores across all disease classes, demonstrating its robustness and effectiveness.

$$\text{Confusion Matrix: } \begin{bmatrix} 18 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 30 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 24 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 26 & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 22 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 & 34 \end{bmatrix}$$

The confusion matrix above provides information on how the model has performed in all of its 42 classes. In the matrix, the rows represent the true classes and the columns represent the predicted classes. Diagonal elements are correct predictions of class, while non-diagonal elements represent misclassifications, indicating instances that have been wrongly predicted by the model.

In this case, the confusion matrix shows that the model performs exceptionally well with all its predictions falling perfectly along the diagonal. This implies that the model classified each and every instance for both the diseases correctly, indicating no errors in handling the dataset.

5 Conclusion

This study demonstrates the usability of machine learning in disease classification based on symptom descriptions. The integration of TF-IDF and LinearSVC achieves high accuracy, while supplementary data, such as disease descriptions and precautions, increase the system's practical applicability. Future work may involve extending the dataset or using more advanced models to cope with evolving health requirements.

References

1. Li, Y., Han, Y., Zhang, X. (2018). *Text Mining Techniques in Health Data Analysis: A Review.* Journal of Medical Informatics, 45(2), 120-130.
2. Karystianis, G., Nevado-Holgado, A., Kim, C. H. (2019). *Applications of TF-IDF in Clinical Text Analysis.* Computational Biology and Medicine, 103(3), 34-42.
3. Akhtar, M., Alam, T., Khan, M. (2020). *The Role of SVM in Disease Diagnosis: Applications and Challenges.* Advances in Machine Learning Applications, 12(1), 89-102.
4. Sharma, R., Gupta, S., Joshi, V. (2019). *Machine Learning Models for Infectious Disease Classification.* International Journal of Healthcare Informatics, 8(4), 56-62.
5. Zhang, L., Wang, Y., Zhu, J. (2021). *Augmenting AI-based Healthcare Systems with Domain Knowledge.* Journal of Intelligent Systems, 17(2), 230-245.

6. Joachims, T. (1998). *Text Categorization with Support Vector Machines: Learning with Many Relevant Features.* Proceedings of the 10th European Conference on Machine Learning (ECML), 137-142.
7. Cortes, C., Vapnik, V. (1995). *Support Vector Networks.* Machine Learning, 20(3), 273-297.
8. Salton, G., Buckley, C. (1988). *Term-weighting Approaches in Automatic Text Retrieval.* Information Processing & Management, 24(5), 513-523.
9. Aggarwal, C. C., Zhai, C. (2012). *A Survey of Text Classification Algorithms.* Mining Text Data, Springer, 163-222.
10. Liu, J., Sun, J., Gao, X. (2019). *Artificial Intelligence in Healthcare: Past, Present, and Future.* Health Informatics Journal, 25(2), 611-622.
11. Rajkomar, A., Dean, J., Kohane, I. (2019). *Machine Learning in Medicine.* New England Journal of Medicine, 380(14), 1347-1358.
12. Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space.* arXiv preprint arXiv:1301.3781.
13. Reddy, S., Fox, J., Purohit, M. P. (2019). *Artificial Intelligence-enabled Healthcare Delivery.* Journal of the Royal Society of Medicine, 112(1), 22-28.
14. Tian, Y., Zhang, Y., Wu, J. (2017). *An Overview of Text Mining in Medicine and Healthcare.* Journal of Biomedical Informatics, 76, 9-27.
15. Zhou, X., He, J., Wu, Z. (2020). *Applications of Support Vector Machines in Biomedical Text Mining.* ACM Transactions on Computing for Healthcare, 1(2), 1-15.