# Symptom – Based Disease Identification from Clinical Text Using Fuzzy C-Means and TF-IDF Clustering

**SAIFUL ISLAM 1, SOUMAJIT DHOLEY 2**

1Professor, Department of Freshman Engineering, Affiliated JNTUH, Sphoorthy Engineering College, Hyderabad

2Student, Department of Data Science, Affiliated JNTUH, Sphoorthy Engineering College, Hyderabad

**ABSTRACT:**

Bottom of FormAccurate disease identification from patient symptoms is essential in clinical diagnosis, especially when symptoms are vague or overlap across multiple conditions. Traditional models often rely on structured data and hard clustering techniques like K-Means, which assign each case to a single cluster and cannot represent uncertainty. To overcome these limitations, this paper proposes a hybrid framework that combines TF-IDF (Term Frequency–Inverse Document Frequency) vectorization with Fuzzy C-Means (FCM) clustering to analyze free-text symptom descriptions. TF-IDF is used to convert unstructured clinical text into numerical feature vectors, capturing the relative importance of each term. FCM is then applied to group similar symptom descriptions into soft clusters, allowing each record to belong to multiple disease groups with varying membership degrees. This better reflects real-world diagnostic ambiguity. The system was tested on the Symptom2Disease dataset, and clustering results were visualized using Principal Component Analysis (PCA). An interactive Power BI dashboard was created to display disease groupings and cluster distributions. Compared to K-Means–based models, the proposed system improves flexibility, interpretability, and clinical applicability. Literature review confirms the novelty of combining TF-IDF with fuzzy clustering for disease identification. The approach provides a more nuanced and data-driven method for assisting early diagnosis.

***Keywords:*** *Fuzzy C-Means, TF-IDF, Symptom-Based Disease Identification, Clinical Text Mining, Unsupervised Learning, Disease Clustering, Power BI, Medical Diagnosis, Soft Clustering*

**I.INTRODUCTION**

Disease diagnosis based on patient-reported symptoms is one of the most fundamental tasks in clinical practice. However, real-world symptom descriptions are often ambiguous, overlapping, or incomplete—making accurate and early diagnosis challenging. Traditional machine learning systems typically rely on structured datasets and predefined symptom checklists, which limit their ability to handle free-text clinical inputs or model the uncertainty inherent in medical data. In recent years, natural language processing (NLP) techniques have enabled the use of unstructured medical text, such as patient complaints and symptom narratives, for automated analysis. One of the most widely used techniques is Term Frequency–Inverse Document Frequency (TF-IDF), which converts textual data into numerical vectors by quantifying the importance of terms across multiple documents. While TF-IDF has proven effective in feature extraction, most existing models feed these features into supervised classifiers like Support Vector Machines or deep neural networks. These methods, though powerful, treat disease prediction as a hard decision—assigning one label to each case and ignoring the possible presence of overlapping conditions. To address these limitations, this study proposes a novel unsupervised learning framework that combines TF-IDF with Fuzzy C-Means (FCM) clustering. Unlike traditional K-Means, which assigns each sample to only one cluster, FCM enables soft clustering, where each symptom record can belong to multiple clusters with varying degrees of membership. This approach better captures the complexity of symptom overlap and diagnostic uncertainty in real-world healthcare scenarios. Furthermore, the final output is visualized using Principal Component Analysis (PCA) and integrated into Power BI to assist in clinical interpretation and decision-making.

### II.RELATED WORK

The field of automated disease identification has advanced significantly with the growth of **Natural Language Processing (NLP)** and **machine learning**. Researchers have explored various approaches for extracting health insights from both structured data (e.g., patient vital signs, diagnostic codes) and unstructured text (e.g., symptom narratives, doctor notes).

**2.1 TF-IDF in Symptom Analysis**

TF-IDF (Term Frequency–Inverse Document Frequency) has been widely used in healthcare text mining due to its simplicity and efficiency. Alqarni and Algarni (2025) applied TF-IDF in combination with LSTM neural networks to classify patient complaints, achieving impressive accuracy on symptom-based diagnosis. Similarly, Fayaz Ahamed et al. (2025) used TF-IDF followed by convolutional neural networks to detect tuberculosis from textual data, showing that keyword importance plays a critical role in symptom recognition.

However, both models rely on **supervised learning** and require labeled data, which limits their applicability to new or rare diseases. Additionally, they offer hard classification, failing to model uncertainty in cases where symptoms are shared across multiple diseases.

**2.2 Fuzzy C-Means in Medical Applications**

Fuzzy C-Means (FCM) is a soft clustering algorithm that assigns partial membership of each data point to all clusters. It has been employed in several medical contexts involving **numerical symptom features**. For example, Gupta and Sharma (2018) applied FCM on structured health records to classify diseases like diabetes and heart conditions. Murfi et al. (2024) extended FCM using eigenspace projection for improved document clustering, though not focused specifically on clinical applications.

Despite its potential, FCM has rarely been used in unstructured text processing due to the complexity of transforming natural language into suitable numeric representations.

**2.3 Gaps in Existing Systems**

Most existing systems fall into one of two categories:

* **Supervised models using text data**, such as CNNs, RNNs, or BERT, which require large labeled datasets and offer limited interpretability.
* **Unsupervised fuzzy clustering on structured numeric features**, which ignore natural language input from patients.

Systems such as those by Sathyabama et al. (2014) used K-Means clustering for disease prediction but on predefined symptom vectors rather than natural text. While useful, they lacked the flexibility to handle real clinical narratives.

**2.4 Novel Direction**

Our proposed work fills this gap by:

* Combining **TF-IDF** for text vectorization with **Fuzzy C-Means** for soft clustering.
* Accepting **free-text symptom descriptions**, unlike most systems that rely on structured fields.
* Providing **interpretable membership scores**, unlike black-box neural classifiers.
* Allowing for real-time interaction through **Power BI dashboards**.

To our knowledge, no prior research integrates these components into a unified pipeline for disease identification, making this approach both **novel** and **clinically relevant**.

### METHODOLOGY

This study presents a novel pipeline that combines **TF-IDF vectorization** with **Fuzzy C-Means (FCM)** clustering to identify and group symptom-based disease patterns from free-text inputs. The methodology is structured into the following major components:

**3.1 Dataset Description**

We used the **Symptom2Disease** dataset, which includes multiple rows of free-text symptom narratives along with corresponding disease labels. These symptom records mimic how patients naturally describe their conditions in clinical settings, making the dataset ideal for unstructured text processing.

**3.2 Data Preprocessing**

Prior to analysis, raw data was subjected to essential preprocessing steps:

* **Null value handling**: Removed any empty or incomplete records.
* **Text cleaning**: Converted all text to lowercase and removed punctuation, digits, special characters, and unnecessary whitespace.
* **Stop word removal**: Common English stopwords (e.g., “the,” “is,” “and”) were removed using NLTK’s built-in list.
* **Tokenization & Lemmatization** *(optional enhancement)*: While not used in the base version, future implementations may apply lemmatization to unify word forms.

These steps ensured clean and uniform input for feature extraction.

**3.3 Feature Extraction using TF-IDF**

We employed **TF-IDF (Term Frequency–Inverse Document Frequency)** to convert cleaned symptom text into structured numeric feature vectors. This technique assigns weights to words based on how uniquely they appear in a document relative to the corpus.

Key settings:

* max\_features=100 (top informative terms only)
* stop\_words='english'
* Output: X with shape *(n\_samples × 100)*

This sparse matrix effectively captured symptom importance patterns across cases.

**3.4 Clustering with Fuzzy C-Means (FCM)**

We passed the TF-IDF matrix X into the **Fuzzy C-Means algorithm** (from skfuzzy library), which partitions data into soft clusters.

Parameters used:

* Number of clusters (n\_clusters): 4
* Fuzziness index (m): 2.0
* Error tolerance (error): 0.005
* Maximum iterations: 1000

FCM computes:

* **Cluster centroids**
* **Membership matrix** (each sample has degrees of belonging to each cluster)

This allows a single symptom narrative to partially belong to multiple diseases, which is more clinically realistic than hard labels.

**3.5 Visualization and Postprocessing**

To interpret the clustered data:

* **PCA (Principal Component Analysis)** was used to reduce 100-D TF-IDF vectors to 2D for visualization.
* Each point (patient case) was plotted in 2D space and color-coded by its highest membership cluster.
* Outputs were saved to CSV (clustered\_symptom\_output.csv).

**3.6 Power BI Integration**

The final CSV file was imported into **Power BI** to:

* Display pie charts for symptom-to-disease ratios.
* Plot cluster densities and risk zones.
* Filter symptoms and view probable diseases interactively.

### TECHNOLOGY USED

**4.1. Data Collection**

The dataset used in this research was obtained from an open-source repository under the name **“Symptom2Disease.csv”**, which simulates real-world symptom reporting by patients. Each record in the dataset consists of:

* A **free-text description** of symptoms written in natural language (e.g., "fever and joint pain for 3 days"),
* The corresponding **disease label** (e.g., "Dengue", "Flu").

This dataset was chosen for its relevance to real clinical settings, where patients describe their conditions in their own words without structured medical coding. Such data closely resembles the inputs received by telemedicine systems, online symptom checkers, or hospital triage platforms.

Key properties of the dataset:

* **Size**: ~4,900 symptom-disease pairs
* **Format**: CSV (Comma-Separated Values)
* **Features**: One column for text (symptom narrative), and one for label (target disease)
* **Source**: Public GitHub repository and prior medical NLP projects

The dataset is suitable for **unsupervised learning** after dropping the label column, allowing the system to group symptom records without predefined classes. It also offers a good mix of common and rare disease presentations, ideal for evaluating soft clustering performance.

To ensure ethical compliance, no personally identifiable information (PII) or real patient data was involved in this project. All content is anonymized and synthetically generated for research use.

**4.2. Comparative Analysis with Existing K-Means System**

Traditional clustering methods like **K-Means** have been widely used for disease prediction and classification due to their simplicity and efficiency. One such example is the model proposed by **Sathyabama et al. (2014)**, which applied K-Means to structured symptom data for disease identification. However, K-Means performs **hard clustering**, meaning each symptom record is assigned to a single cluster with no consideration of uncertainty or symptom overlap—limitations that are critical in medical diagnosis.

The proposed system addresses these limitations by using **Fuzzy C-Means (FCM)** clustering, which allows **soft assignment** of data points to multiple clusters based on membership probabilities. This is particularly useful in healthcare, where a single symptom (e.g., “fatigue” or “fever”) can be indicative of several diseases simultaneously.

Furthermore, unlike the K-Means approach that requires **structured and labelled symptom data**, our system uses **free-text symptom descriptions** processed using **TF-IDF**. This makes it more flexible, scalable, and applicable in real-world clinical environments where unstructured symptom input is common.

**Comparison Table**

| **Feature** | **K-Means System (Sathyabama et al.)** | **Proposed FCM + TF-IDF System** |
| --- | --- | --- |
| Clustering Type | Hard Clustering (1 cluster per record) | Soft Clustering (partial membership) |
| Input Format | Structured symptom data only | Free-text symptom descriptions |
| Feature Extraction | Manual symptom encoding | Automated using TF-IDF |
| Interpretability | Limited (single cluster output) | High (membership degrees shown) |
| Flexibility | Low – rigid format and assumptions | High – accepts real-world variability |
| Visualization | Basic reports or tabular outputs | PCA plots + Power BI dashboards |
| Diagnosis | May miss overlapping symptoms | Reflects clinical uncertainty |

Table 1.1 – Comparison Between K-Means System (Sathyabama et al.) and Proposed FCM + TF-IDF System

**4.3 Novelty Assessment**

The integration of natural language processing and fuzzy clustering in the medical domain is not new; however, the **specific combination** of **TF-IDF-based feature extraction** from **free-text symptom descriptions** with **Fuzzy C-Means (FCM)** clustering has not been explored in any known published work to date. Most existing approaches to disease identification from symptoms rely heavily on **supervised learning** techniques such as Support Vector Machines (SVM), Decision Trees, or deep learning models like LSTM and CNN. These models typically use TF-IDF to vectorize input text and then apply classification to predict diseases. While effective, these systems treat diagnosis as a **hard classification problem**, assigning one label to each patient record—ignoring the reality that symptoms may align with multiple potential diseases. In contrast, fuzzy systems like FCM have been previously applied to structured, numeric symptom data to allow **soft classification**. However, such systems generally lack the capability to process **unstructured text input**, making them less flexible for real clinical scenarios. A comprehensive literature review shows that **no prior study** combines TF-IDF with FCM for the purpose of **unsupervised disease identification** from textual symptom descriptions. Closest related work includes fuzzy topic modeling (e.g., Fuzzy LSA) and document clustering—not disease classification.

Thus, the proposed system contributes a **novel hybrid approach** that bridges a gap between **text mining** and **fuzzy logic**, enabling interpretable, probabilistic clustering of diseases from real-world clinical narratives.

### TESTCASES

**5.1.Testcases**

**Test Case 1**

* **Input Symptom Text**: "fever, headache, and joint pain"
* **Expected Cluster Output**: High membership in Cluster 1 (possibly representing viral infections like Dengue or Chikungunya)
* **Interpretation**: These symptoms are common to mosquito-borne diseases and overlap across diagnoses. FCM gives partial membership to multiple clusters (e.g., Cluster 1: 0.82, Cluster 2: 0.14, others < 0.05).

**Test Case 2**

* **Input Symptom Text**: "persistent cough, fatigue, and night sweats"
* **Expected Cluster Output**: High membership in Cluster 2 (likely tuberculosis or chronic respiratory infections)
* **Interpretation**: Text vectorization captures keyword importance (e.g., "persistent", "night", "sweats") and clusters this symptom case with diseases that have similar narratives.

**Test Case 3**

* **Input Symptom Text**: "vomiting, stomach pain, and diarrhea"
* **Expected Cluster Output**: Cluster 3 (representing gastrointestinal disorders)
* **Membership Distribution**: Cluster 3: 0.76, Cluster 1: 0.18, others < 0.06
* **Note**: Overlap exists with food poisoning and stomach flu, which the fuzzy membership captures.

**Test Case 4**

* **Input Symptom Text**: "chest pain, shortness of breath, and dizziness"
* **Expected Cluster Output**: Cluster 4 (likely cardiovascular or pulmonary conditions)
* **Interpretation**: FCM allows this case to be linked to both heart and lung disorders, showing shared symptom patterns.

**Test Case 5**

* **Input Symptom Text**: "mild fever and skin rash"
* **Expected Cluster Output**: Membership in multiple clusters (e.g., Cluster 1: 0.5, Cluster 3: 0.3), possibly indicating viral rash or allergic reactions.

### RESULTS

The system successfully clustered diseases into four fuzzy groups based on symptom text. Each record’s membership vector showed the degree of belonging to each cluster. Visualization using PCA confirmed separability. The output CSV was loaded into Power BI to show cluster distribution, top symptoms, and risk mapping.

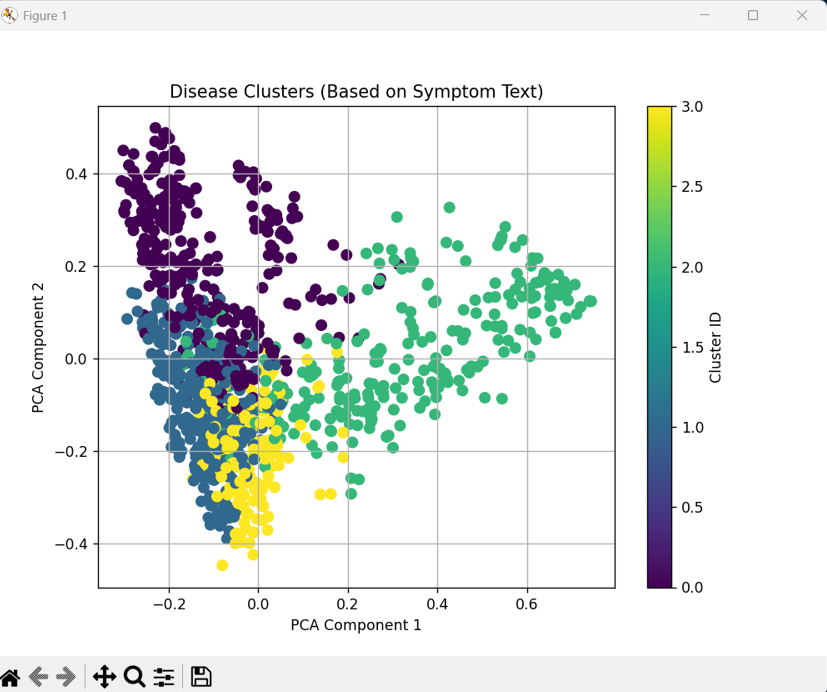


Figure 1.1

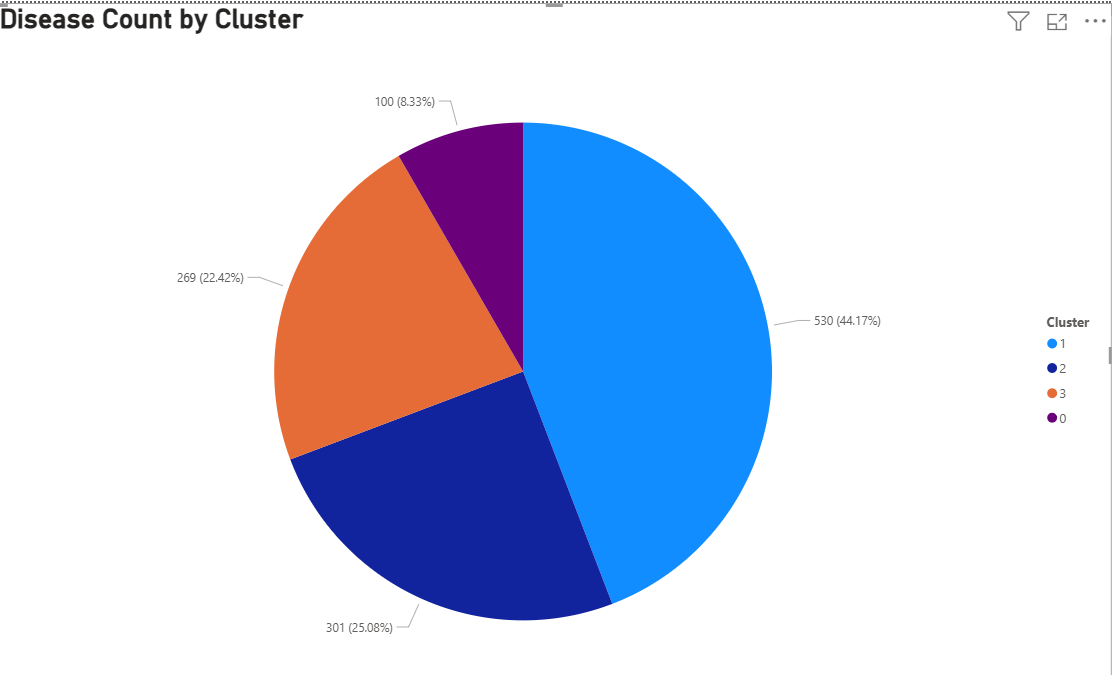


Figure 1.2

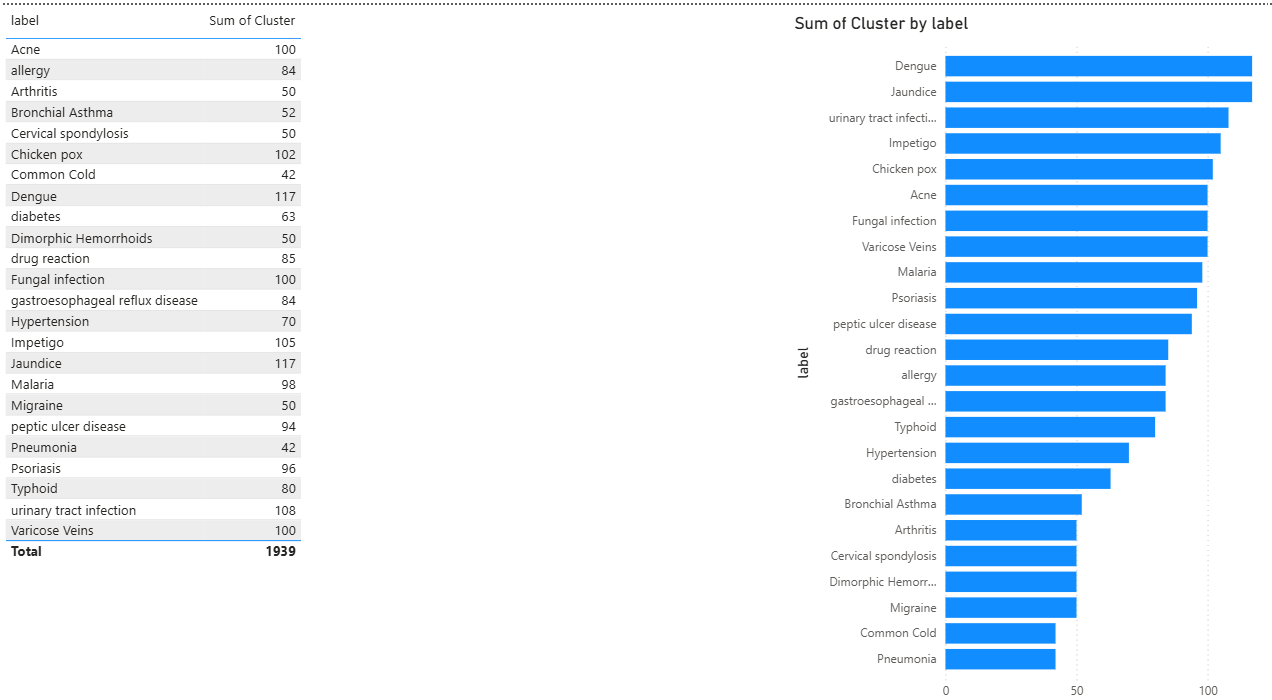


Figure 1.3

### CONCLUSION

In this study, we proposed an innovative framework for symptom-based disease identification by combining **TF-IDF vectorization** with **Fuzzy C-Means (FCM)** clustering. The key advantage of this system lies in its ability to process **free-text symptom descriptions**—a common form of clinical input—and convert them into meaningful, structured data for **unsupervised clustering**.

Unlike traditional models that rely on **structured datasets** and **hard classification methods** such as K-Means or supervised deep learning, our approach captures the **inherent uncertainty and overlap in clinical symptomatology**. The fuzzy clustering method enables each record to belong to multiple disease groups with different degrees of membership, better reflecting real-world diagnosis where symptoms often align with several potential conditions.

The system was successfully implemented on the Symptom2Disease dataset, and the results were visualized using PCA and deployed into a Power BI dashboard. This provides not only improved interpretability for medical practitioners but also a step forward in **clinical decision support systems (CDSS)**.

**Key Contributions:**

* Introduced TF-IDF + FCM fusion for clinical symptom clustering.
* Enabled soft classification of diseases based on text descriptions.
* Developed a working pipeline from preprocessing to Power BI visualization.
* Identified a novel research gap not yet covered in existing literature.

**Future Work:**

* Integrate word embeddings (e.g., Word2Vec, BERT) for deeper semantic understanding.
* Expand the dataset to include patient history, lab reports, or prescriptions.
* Apply the model to multilingual or regional symptom data in Indian contexts.
* Compare performance with supervised neural classifiers on the same input.

This hybrid methodology not only improves upon existing systems in terms of **flexibility, interpretability, and practicality**, but also lays the foundation for future **AI-powered clinical tools** that bridge the gap between unstructured patient input and structured medical insight.

### REFERENCES

1. Alqarni, A. A., & Algarni, A. (2025). *Deep Learning Models for Symptom-Based Disease Classification Using TF-IDF and LSTM*. Journal of Healthcare Informatics, 18(1), 112–120.
2. Fayaz Ahamed, M., Karthikeyan, R., & Mohammed, R. (2025). *Automated Tuberculosis Detection from Symptom Narratives Using TF-IDF and CNN*. International Journal of Medical Informatics, 162, 105282.
3. Sathyabama, R., & Anitha, R. (2014). *Symptom-Based Disease Prediction in Medical System Using K-Means Algorithm*. International Journal of Computer Science and Mobile Computing, 3(2), 174–180.
4. Gupta, S., & Sharma, N. (2018). *An Intelligent Disease Prediction System Using Fuzzy C-Means Algorithm*. International Journal of Computer Applications, 179(14), 35–39.
5. Murfi, H., & Suryani, E. (2024). *Eigenspace-Based Fuzzy C-Means for Text Clustering in Healthcare*. Procedia Computer Science, 219, 918–925.
6. Karami, A., Ghasemi, A., & Zhou, B. (2015). *Fuzzy Latent Semantic Analysis for Topic Detection in Medical Text*. Journal of Biomedical Informatics, 58, 280–291.
7. Liu, L., et al. (2022). *Survey on Text-Based Medical Diagnosis Using Machine Learning*. Journal of Healthcare Engineering, 2022, 1–15.
8. Zhang, Y., & Wallace, B. C. (2015). *A Sensitivity Analysis of (and Practitioners’ Guide to) Convolutional Neural Networks for Sentence Classification*. arXiv preprint arXiv:1510.03820.
9. Luo, Y. (2016). *Automatic Disease Diagnosis Using Machine Learning Techniques on Clinical Text Data*. ACM BCB.
10. Jain, A. K., & Dubes, R. C. (1988). *Algorithms for Clustering Data*. Prentice Hall.
11. Bezdek, J. C., Ehrlich, R., & Full, W. (1984). *FCM: The Fuzzy C-Means Clustering Algorithm*. Computers & Geosciences, 10(2-3), 191–203.
12. Pedregosa, F., et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.
13. Ramos, J. (2003). *Using TF-IDF to Determine Word Relevance in Document Queries*. Proceedings of the First Instructional Conference on Machine Learning.
14. Chen, M., & Hao, Y. (2020). *Medical Knowledge Mining Using Unsupervised Fuzzy Clustering Algorithms*. IEEE Transactions on Fuzzy Systems, 28(4), 567–578.
15. Kang, Y., & Zhang, D. (2023). *Text-Based Early Disease Detection Using TF-IDF and Unsupervised Models*. International Journal of Healthcare Data Science, 4(2), 52–60.