**Project – 3**

**Development of a Fuzzy Logic Disease Prediction System**

**1. Introduction and problem overview**

Respiratory illnesses such as COVID-19, influenza (flu), the common cold, and allergies often present with overlapping symptoms such as cough, fever, and fatigue. These similarities can result in diagnostic delays, inappropriate treatment, and additional healthcare burden. During health crises like the COVID-19 pandemic, the ability to quickly and accurately assess symptoms and suggest probable illnesses became especially critical. Traditional diagnostic methods rely heavily on clinical testing and expert evaluation, which may not always be accessible or timely.

This project addresses the problem by developing an intelligent, interactive system that leverages both fuzzy logic and machine learning (ML) techniques. The goal is to accurately classify respiratory conditions based on patient-reported symptoms using a hybrid model that incorporates human-like reasoning with statistical pattern recognition. The system is designed to be simple, explainable, and capable of functioning even with incomplete or uncertain symptom inputs.

**2. Literature review**

Rule-based expert systems and statistical classifiers have been applied to healthcare problems for decades. For example, symptom checkers using decision trees, support vector machines (SVMs), and Bayesian models are commonly used to assess illnesses from symptom inputs. Although these models offer high accuracy in some cases, they often lack the interpretability and adaptability needed in dynamic, real-world medical contexts.

Fuzzy logic, introduced by Zadeh (1965), is particularly suitable for handling vague or imprecise information, making it ideal for modeling human-like reasoning in symptom evaluation. Prior studies have shown that fuzzy systems can effectively model triage systems, pain assessment, and chronic disease management.

The dataset used in this project, sourced from Kaggle labeled data for COVID-19, flu, cold, and allergy based on 20 binary symptoms. It provides a comprehensive platform for training both ML classifiers and a fuzzy inference system.

**3. Methodology**

**3.1 Dataset Description**

The dataset includes:

* 20 binary features representing symptom presence (e.g., Fever, Cough, Fatigue, Shortness of Breath)
* Target label: TYPE indicating disease category (COVID-19, FLU, COLD, ALLERGY)
* Over 12,000 data records after cleaning and preprocessing

This dataset was preprocessed to remove null values and irrelevant columns. The symptoms were encoded as 0 (no) or 1 (yes), and the disease type was used as the target variable.

**3.2 Fuzzy Logic System Design**

Fuzzy input variables were created for each binary symptom with two linguistic terms: low and high, corresponding to absence or presence. The fuzzy output variable risk was defined with linguistic labels: low, medium, and high, spanning a universe from 0 to 10.

Rules were generated using combinations of symptom presence, ranging from single symptoms to combinations of 3-4. For example:

Rule: If fever and cough the risk is medium

Rule: If fever and cough and fatigue then risk is high

The fuzzy inference engine was built using scikit-fuzzy, and defuzzification was used to produce a risk score between 0–10. This score was then incorporated into final disease predictions.

**3.3 Machine Learning Classification Models**

The following classifiers were implemented using scikit-learn:

* Decision Tree: A shallow tree with max\_depth=5 to maintain interpretability
* Naive Bayes: A probabilistic model assuming independence between symptoms
* Random Forest: An ensemble of 100 decision trees to improve generalization.

Each model was trained on 80% of the dataset and tested on the remaining 20%. Metrics such as accuracy, precision, recall, and F1-score were calculated to compare their performance.

**3.4 Hybrid Disease Prediction Algorithm**

A hybrid algorithm was implemented as follows:

1. User inputs yes/no for 20 symptoms
2. Fuzzy system computes a risk score based on combinations of symptoms
3. For each disease, the user input is compared with average disease symptom profiles
4. A final score is computed for each disease: disease\_score = symptom\_match\_score + normalized\_risk\_score
5. The disease with the highest total score is selected as the most likely diagnosis

Most likely condition: flu [Score: 4.5]

**4. Results**

**4.1 ML Model Performance Summary**

The table below summarizes the classification performance of the models:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Parameters Used** | **Accuracy** | **F1-Score** | **AUC (ROC)** | **Notes** |
| Decision Tree | max\_depth=5, random\_state=42 | 0.923 | 0.4804 | 0.99 | Simple and interpretable. High accuracy, but low F1-score suggests class imbalance bias. |
| Naive Bayes | GaussianNB() (no hyperparameters set manually) | 0.9342 | 0.815 | 0.98 | Best performance overall. High generalization across classes due to probabilistic nature. |
| Random Forest | n\_estimators=100, random\_state=42 | 0.8858 | 0.5723 | 0.99 | Robust ensemble model. Good AUC, moderate balance. Prone to overfitting dominant classes. |
| Fuzzy Logic | Rule-based scoring with symptom-weighted risk normalization | 0.4185 | 0.369 | 0.71 | Lowest performance in metrics but high interpretability and resilience to incomplete input. |

Naive Bayes achieved the highest performance, with an accuracy of 0.9342 and a macro F1-score of 0.815, indicating strong and balanced classification across all disease categories.

Decision Tree showed good accuracy (0.923) but a relatively low F1-score (0.4804), suggesting it may be biased toward majority classes.

Random Forest had decent accuracy (0.8858) and a moderate F1-score (0.5723), performing better than Decision Tree but less balanced than Naive Bayes.

Fuzzy Logic system had the lowest performance, with an accuracy of 0.4185 and F1-score of 0.369, due to its reliance on static, rule-based scoring without adaptive learning.

The results highlight that while fuzzy logic offers interpretability and flexibility, machine learning models outperform it significantly in predictive accuracy and balance across classes.

**4.2 ROC Curve**

The ROC curve shows that machine learning models significantly outperform the fuzzy logic system in classification accuracy. Random Forest and Decision Tree both achieve an AUC of 0.99, and Naive Bayes follows closely with 0.98, indicating excellent performance. In contrast, the fuzzy logic system, with an AUC of 0.71, demonstrates moderate predictive ability. While less accurate, the fuzzy system offers valuable interpretability and handles uncertain or incomplete inputs better than ML models.

A graph of a curve

AI-generated content may be incorrect.

**4.3 sample scenario**

**Case study 1:**

A screenshot of a computer

AI-generated content may be incorrect.

A graph of a graph with lines and numbers

AI-generated content may be incorrect.

**Case study 2:**

**A screenshot of a computer screen

AI-generated content may be incorrect.**

A graph of a triangle with blue and green lines

AI-generated content may be incorrect.

**5. Discussion**

The performance evaluation highlights that Naive Bayes achieved the highest macro-averaged precision and recall, likely due to its probabilistic nature and ability to handle overlapping symptom patterns effectively. Although Random Forest demonstrated strong overall accuracy and the highest AUC in the ROC curve comparison (0.99), it showed signs of class imbalance, tending to favor predictions for flu over less represented classes like COVID and cold. The Decision Tree model, while interpretable and straightforward, showed comparatively lower recall and was limited by its shallow depth, which affected its ability to generalize across diverse symptom profiles.

The fuzzy logic component added significant value by incorporating interpretability and resilience to incomplete or imprecise input. Its ability to model uncertainty using linguistic terms (e.g., "low", "medium", "high" risk) enabled intuitive reasoning, especially in ambiguous cases. When combined with a data-driven profile matching system, the hybrid scoring mechanism enhanced diagnostic reliability and adaptability. This integration of fuzzy logic with machine learning models demonstrates that combining symbolic and statistical approaches can improve performance and interpretability in multi-class, symptom-based disease prediction tasks.

**6. Conclusion**

This project demonstrated a practical, explainable hybrid AI system for symptom-based respiratory illness prediction. The use of fuzzy logic alongside machine learning enables effective classification while maintaining interpretability. The system is user-friendly, modular, and extensible.

**7. Limitations**

The dataset contains only binary symptom inputs (yes/no), lacking severity levels or time-series symptom progression.

The model performance is limited by class imbalance; flu cases dominate, affecting model generalization to rarer classes like COVID.

Fuzzy logic rules are manually constructed and not dynamically learned, which may restrict scalability.

Demographic factors such as age and gender are underutilized and could enhance predictions if integrated properly.

**8. Future directions**

Add symptom severity levels (e.g., mild/moderate/severe)

Integrate time-series symptom tracking

Deploy as a mobile/telehealth application

**9. References**

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control, 8*(3), 338–353. <https://doi.org/10.1016/S0019-9958(65)90241-X>

Conway, W. (n.d.). *COVID, Flu, and Cold Symptoms Dataset* [Data set]. Kaggle. <https://www.kaggle.com/datasets/walterconway/covid-flu-cold-symptoms>

The scikit-fuzzy development team. (n.d.). *scikit-fuzzy documentation*. <https://pythonhosted.org/scikit-fuzzy/>

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830. <https://scikit-learn.org/>