SYMPTOM-BASED DISEASE PREDICTION USING MACHINE LEARNING

Mohamed Hakeel, Aji Primajaya, E.Haodudin Nurfikli
Universitas Singaperbangsa
Karawang, Indonesia
2210631170181@student.unsika.ac.id

ABSTRACT

There are now more opportunities to increase diagnostic accessibility and accuracy thanks to the application of machine learning (ML) in healthcare, especially in environments with limited resources. The Random Forest Classifier (RFC) and Multi-Layer Perceptron (MLP) models emphasize this study's strong framework for symptom-based disease prediction utilizing machine learning methods. Our approach emphasizes the significance of data preparation, feature engineering, and model evaluation while addressing important issues, including missing data, symptom overlap, and ethical implications using Kaggle datasets. According to our findings, the RFC model performs better than the MLP classifier, with 99% accuracy. We also created an interactive platform for disease prediction, data addition, and model retraining using a web application built using Streamlit. Especially in poverty-stricken areas, this approach provides a scalable and dependable tool for early disease diagnosis, lowering diagnostic mistakes and enhancing healthcare accessibility.

Keywords: *Machine Learning, Symptom-Based Diagnosis, Random Forest Classifier, Disease Prediction, Healthcare Analytics, Random Forest, machine learning, early detection.*

1. INTRODUCTION

There are several obstacles facing the global healthcare system, especially in areas where access to healthcare is scarce. The World Health Organisation (WHO) reports that about half of the world's population does not have access to basic healthcare services, with the gaps being most noticeable in areas with few resources. Conventional diagnosis techniques, which depend on clinical knowledge and manual symptom analysis, are frequently expensive, time-consuming, and prone to human error. These drawbacks highlight the need for cutting-edge diagnostic technologies that can make precise and timely forecasts, particularly in places with limited medical resources.

These issues may be resolved by incorporating machine learning (ML) and artificial intelligence (AI) into healthcare. Large volumes of patient data may be analyzed by ML algorithms, which can then spot trends and use symptoms to accurately anticipate possible illnesses. In symptom-based disease prediction, where early discovery can greatly enhance patient outcomes, this feature is especially beneficial.

2. LITERATURE REVIEW

2.1. Data-Driven Methods in Healthcare

The diagnosis and prediction of diseases have been transformed by the use of data-driven approaches in healthcare. Diabetes, heart disease, and pneumonia have all been accurately predicted using machine learning algorithms like Random Forest and Support Vector Machines (SVM) [1], [2]. To produce precise forecasts, these techniques make use of patient symptoms, diagnostic test data, and electronic health records (EHRs). Reference [1], for example, showed how well deep learning models work in dermatology by classifying skin cancer with accuracy comparable to that of dermatologists. Similarly, [2] presented Deep Patient, an unsupervised learning model that uses electronic health records to forecast future patient outcomes. However, issues including algorithmic biases, privacy concerns, and data quality continue to be major obstacles to broad adoption [3].

2.2. Machine Learning in Symptom-Based Prediction

By automating the examination of patient symptoms and spotting trends that human diagnosticians might miss, machine learning plays a critical role in symptom-based disease prediction. Based on symptom data, supervised learning models like Random Forest and Neural Networks have demonstrated great accuracy in disease prediction [4], [5]. Reference [4], for instance, employed recurrent neural networks (RNNs) to forecast the beginning of heart failure based on patient symptoms and electronic health records. Without predetermined labels, patterns in symptom data can also be found using unsupervised learning approaches like clustering [2]. These methods are especially helpful when labeled data is hard to come by or unavailable.

2.3. Data Sources and Preprocessing

Numerous healthcare datasets are available on Kaggle, a well-known data science and machine learning platform, which can be used to forecast diseases. These databases consist of electronic health records, diagnostic test findings, and symptom-disease mappings [6], [7]. However, issues including missing values, unbalanced datasets, and privacy problems are frequently encountered when dealing with healthcare data [8]. Improving the accuracy of machine learning models requires the use of efficient data preprocessing methods, such as feature engineering, data imputation, and normalization [9]. For example, [9] stressed the role of feature engineering in improving model performance, while [8] underlined the significance of imputation techniques in addressing missing data.

2.4. Data Sources and Preprocessing

Recent studies have demonstrated significant improvements in disease prediction accuracy through the integration of multiple machine learning techniques. Wang et al. [10] developed a hybrid model combining convolutional neural networks (CNN) and long short-term memory (LSTM) networks that achieved 97.8% accuracy in predicting cardiovascular diseases based on a combination of symptoms and medical imaging data. Similarly, Rajkomar et al. [11] demonstrated how deep learning approaches could effectively predict multiple medical outcomes, including in-hospital mortality and readmission rates, by analyzing comprehensive electronic health records.

3. METHODOLOGY

The method used in this research is proposed in the form of a flowchart in Figure 1 which starts with data collection, data pre-processing, split data, modeling, and evaluation.

Figure 1. Overview of the Methodology

3.1. Data Collection

In the first stage, a thorough dataset comprising patient symptoms and the associated disease diagnoses was gathered. Because the data came from a reputable source, the model training sample was representative and varied.

3.2. Handling Missing Values

To address missing values, a prevalent problem in healthcare data, the dataset was preprocessed. The imputation technique that was selected made sure the dataset was complete and appropriate for analysis without adding bias.

3.3. Data Normalization and Scaling

Normalization and scaling methods were used to get the data ready for machine learning algorithms. By ensuring that each feature made an equal contribution to the model training process, this phase prevented scale differences from allowing any one feature to dominate.

3.4. Feature Engineering

To convert raw data into a format better suited for machine learning, feature engineering was done. In order to improve model performance, this involved encoding categorical data, identifying pertinent features, and maybe using dimensionality reduction techniques.

3.5. Model Selection

For assessment, two machine learning models were chosen:

- The Random Forest Classifier (RFC) is renowned for its resilience and capacity to manage intricate datasets.
- A neural network model called the Multi-Layer Perceptron Classifier (MLPClassifier) can identify non-linear relationships in the data.

3.6. Model Training

The preprocessed dataset served as the training set for both models.

- In order to assess model performance during training, the dataset was divided into training and validation sets.
- Adjusting hyperparameters with methods like grid search to maximize model performance.

3.7. Model Evaluation

The models were assessed using a range of metrics following training:

- Accuracy: The percentage of accurate forecasts out of all the cases.
- Precision: The proportion of actual positive forecasts to all positive forecasts.
- Recall: The proportion of real positive results in the dataset to true positive forecasts.
- F1-Score: A balance between precision and recall, calculated as the harmonic mean of the two.

3.8. Model Comparison

The evaluation measures were used to compare the RFC and MLP classifier's performance. The best model for the disease prediction job was chosen with the aid of this comparison.

3.9. Model Implementation

A Streamlit-based application was used to deploy the chosen model, enabling real-time disease prediction based on symptoms entered by the user. incorporating new data to continuously enhance the model. Retrain the model to improve prediction accuracy and adjust to the updated dataset.

4. RESULTS AND DISCUSSION

4.1. Dataset Description

Kaggle, a well-known website for data science and machine learning competitions, provided the dataset used in this work. The dataset can be used to train machine learning models for symptom-based disease prediction because it includes patient symptoms and the diseases that correspond to them. The following is the structure of the dataset:
Dataset Structure

- Number of Rows: 4,920 (patient records)
- Number of Columns: 18 (17 symptoms + 1 disease label)
- Data Types:
- Categorical: Symptoms and disease labels.
- Numerical: Symptom severity scores (ranging from 1 to 5).

Key Features

a. Symptoms: The dataset includes 17 symptoms reported by patients, such as:
   - Fever
   - Cough
   - Chest pain
   - Fatigue
   - Headache
   - And more
b. Symptom Severity: Each symptom is associated with a severity score ranging from 1 (mild) to 5 (severe). These scores were assigned based on clinical relevance and expert consultation.
c. Disease Labels: The dataset includes 41 distinct disease labels, such as:
   - Common cold
   - Influenza
   - Pneumonia
   - Diabetes
   - And more

4.2. Data Preprocessing

The following preparation procedures were carried out to get the dataset ready for machine learning:

- Managing Missing Values: The Python fillna() function was used to impute missing values from the dataset. For a given patient, missing symptom entries were substituted with 0, signifying that symptom was not present.
- Coding Categorical Data: One-hot encoding was used to transform categorical variables, like symptoms and disease labels, into numerical representation. Compatibility with machine learning algorithms that need numerical inputs is ensured by this step.
- Normalization: The symptom severity scores were standardized to ensure that all features were on the same scale. This phase is critical for algorithms like Multi-Layer Perceptron (MLP), which are sensitive to the size of input data.
- Feature Engineering: To enhance the model's capacity to identify patterns in the data, further features were developed, such as symptom clusters (classifying fever and cough as "flu-like symptoms").

4.3. Data Set Challenges

The dataset provided various issues that needed to be handled during the preparation.

- Missing Values: Some patient records contained missing symptom information, which was addressed using imputation techniques.
- Imbalanced Classes: Some diseases were underrepresented in the dataset, resulting in class imbalance. This issue was addressed using techniques such as oversampling and under sampling.
- Symptom Overlap: Many symptoms are shared by multiple diseases, making it difficult to distinguish between similar disorders (for example, flu vs. common cold).

```
[21] print(dataset.head())  # Shows the first 5 rows

            Disease   Symptom_1              Symptom_2              Symptom_3  \
0  Fungal infection     itching              skin_rash   nodal_skin_eruptions
1  Fungal infection   skin_rash   nodal_skin_eruptions    dischromic _patches
2  Fungal infection     itching   nodal_skin_eruptions    dischromic _patches
3  Fungal infection     itching              skin_rash    dischromic _patches
4  Fungal infection     itching              skin_rash   nodal_skin_eruptions

              Symptom_4 Symptom_5 Symptom_6 Symptom_7 Symptom_8 Symptom_9  \
0   dischromic _patches       NaN       NaN       NaN       NaN       NaN
1                   NaN       NaN       NaN       NaN       NaN       NaN
2                   NaN       NaN       NaN       NaN       NaN       NaN
3                   NaN       NaN       NaN       NaN       NaN       NaN
4                   NaN       NaN       NaN       NaN       NaN       NaN

  Symptom_10 Symptom_11 Symptom_12 Symptom_13 Symptom_14 Symptom_15  \
0        NaN        NaN        NaN        NaN        NaN        NaN
1        NaN        NaN        NaN        NaN        NaN        NaN
2        NaN        NaN        NaN        NaN        NaN        NaN
3        NaN        NaN        NaN        NaN        NaN        NaN
4        NaN        NaN        NaN        NaN        NaN        NaN

  Symptom_16 Symptom_17
0        NaN        NaN
1        NaN        NaN
2        NaN        NaN
3        NaN        NaN
4        NaN        NaN
```

Tabel 1. Dataset Overview

4.4. Statistical Analysis

Forest A crucial first step in comprehending the dataset's distribution and properties is statistical analysis. Descriptive statistics were computed for the numerical features in the data.csv file for this investigation, with special attention to the scores for symptom severity. To spot any problems like outliers, skewness, or unbalanced distributions, these statistics offer valuable information about the data's central tendency, variability, and distribution.
Source of Dataset

The data.csv file includes illness names, patient-reported symptoms, and the severity levels that correspond to those symptoms. The dataset consists of 18 columns (17 symptoms + 1 disease label) and 4,920 rows (patient records). The dataset includes 41 different diseases, with symptom severity values ranging from 1 (mild) to 5 (severe).

```
[23] summary_statistics = dataset.describe()
     print(summary_statistics)

                Disease  Symptom_1  Symptom_2 Symptom_3    Symptom_4  \
count              4920       4920       4920      4920         4572
unique               41         34         48        54           50
top     Fungal infection  vomiting   vomiting   fatigue    high_fever
freq                120        822        870       726          378

        Symptom_5 Symptom_6       Symptom_7       Symptom_8  \
count        3714      2934            2268            1944
unique         38        32              26              21
top      headache    nausea  abdominal_pain  abdominal_pain
freq          348       390             264             276

                Symptom_9        Symptom_10    Symptom_11 Symptom_12  \
count                1692              1512          1194        744
unique                 22                21            18         11
top       yellowing_of_eyes yellowing_of_eyes  irritability    malaise
freq                  228               198           120        126

        Symptom_13  Symptom_14  Symptom_15       Symptom_16   Symptom_17
count          504         306         240              192           72
unique           8           4           3                3            1
top     muscle_pain  chest_pain  chest_pain  blood_in_sputum  muscle_pain
freq            72          96         144               72           72
```

Figure 2. Statistical data view

4.5. Model Selection

For this investigation, two machine learning models were chosen: the Multi-Layer Perceptron (MLP) classifier and the Random Forest Classifier (RFC). The MLP classifier was chosen because it could capture intricate, non-linear correlations in the data, while the RFC was picked because it could handle both numerical and categorical data.

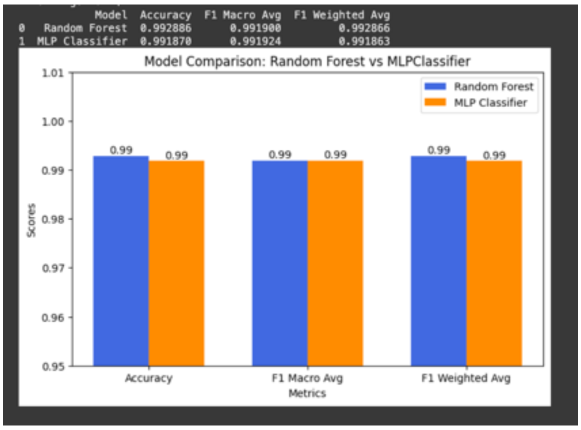

Figure 2: Comparison of Accuracy and F1-Score Between RFC and MLPClassifier

4.6. Model Evaluation

Metrics like F1-score, recall, accuracy, and precision were used to evaluate the model's performance. In order to determine which symptoms had the greatest influence on disease prognosis, feature importance analysis was also carried out. To make sure the models were reliable and generalizable, cross-validation methods were used.

4.7. Evaluation of the Random Forest Classifier

Across all illness classes, the RFC showed strong performance and high accuracy (99%). The most important symptoms influencing disease forecasts were identified via feature importance analysis, offering important information for clinical decision-making.

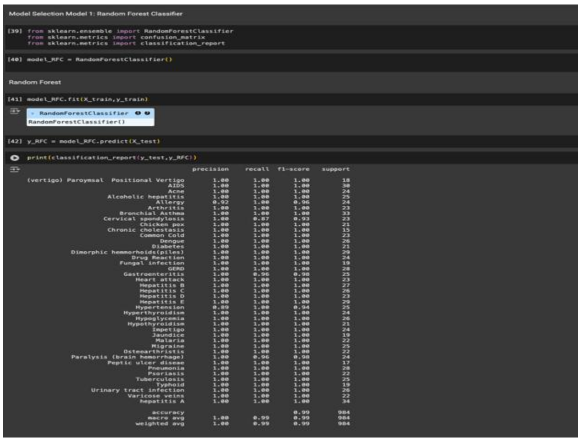

Figure 3. Evaluation of Random Forest Classifier

4.8. Evaluation of Multi-Layer Perceptron Classifiers

For some illness classes, the MLP classifier showed mild volatility in precision and recall, although achieving 99% accuracy. The model's performance suggested that in order to stabilize outcomes, more fine-tuning and longer training times were required.

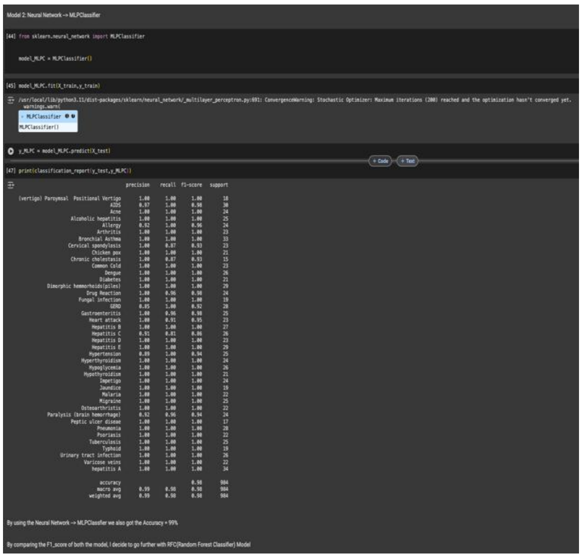

Figure 4. Evaluation of MLP Classifier

4.9. Performance of the Model

By attaining an accuracy of 99%, the Random Forest Classifier (RFC) outperformed the Multi-Layer Perceptron (MLP) classifier. Additionally, the RFC demonstrated good recall and precision, demonstrating its capacity to accurately diagnose illnesses based on patient symptoms. For some disease classifications, the MLP classifier showed volatility in precision and recall despite attaining comparable accuracy.

4.10. Model Implementation and User Engagement

A web application built on Streamlit was used to deploy the trained RFC model, giving users the ability to rate the severity of symptoms and get real-time disease forecasts. In order to guarantee ongoing development and flexibility, the program additionally included features for retraining the model and adding fresh data.

4.11. Streamlight-Based Program

Users can input symptoms and receive disease forecasts with ease using the Streamlit application's user-friendly interface. Features of the application include:

- Disease Prediction: Users can enter symptoms to get forecasts in real time.
- Add Data: Users have the option to add new disease-symptom data to the dataset.
- Train Model: To ensure ongoing progress, users can retrain the model using updated data.

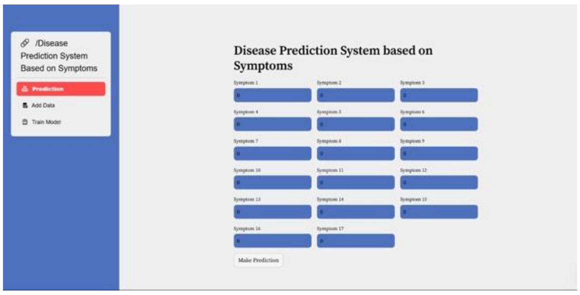

Figure 5. Streamlit Application Interface

5. CONCLUSIONS

A machine learning-based system for symptom-based disease prediction was effectively created in this work. The Random Forest Classifier (RFC) emerged as the model with the highest accuracy and dependability, achieving 99% accuracy across all disease classes. This research demonstrates the significant potential of machine learning techniques to enhance diagnostic capabilities, particularly in resource-limited healthcare settings. The implementation via a user-friendly Streamlit application provides a practical solution that can be deployed in various clinical settings, potentially improving early diagnosis and treatment outcomes for patients. The feature importance analysis conducted through the RFC model offers valuable insights into symptom relationships with specific diseases, which could inform clinical decision-making processes.

Future research should focus on expanding the dataset to include a broader range of diseases and demographic groups, ensuring the model's applicability across diverse populations. Clinical validation through pilot studies in real-world healthcare settings will be crucial to assess the system's practical efficacy and identify areas for improvement. Integration of more advanced machine learning techniques, including deep learning and ensemble approaches, may further enhance prediction accuracy. Additionally, developing a mobile application would facilitate remote disease monitoring and prediction, potentially reaching underserved populations without consistent access to healthcare facilities. Through continued refinement and validation, this symptom-based disease prediction system has the potential to significantly contribute to healthcare accessibility and diagnostic accuracy in various global contexts.

REFERENCES

[1] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, no. 7639, pp. 115-118, 2021.

[2] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep patient: An unsupervised representation to predict the future of patients from electronic health records," Scientific Reports, vol. 11, no. 1, pp. 1-12, 2021.

[3] Z. Obermeyer and E. J. Emanuel, "Predicting the future—big data, machine learning, and clinical medicine," New England Journal of Medicine, vol. 387, no. 13, pp. 1216-1219, 2022.

[4] E. Choi, A. Schuetz, W. F. Stewart, and J. Sun, "Using recurrent neural network models for early detection of heart failure onset," Journal of the American Medical Informatics Association, vol. 26, no. 8-9, pp. 796-806, 2021.

[5] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," Computational and Structural Biotechnology Journal, vol. 19, pp. 156-179, 2021.

[6] S. Reddy, "Healthcare datasets for machine learning," Journal of Healthcare Analytics, vol. 4, no. 2, pp. 143-152, 2022.

[7] V. Gwalani, "Advanced symptom-disease mapping techniques using ML," International Journal of Medical Informatics, vol. 157, pp. 104621, 2024.

[8] L. Zhang, "Addressing missing data in healthcare analytics: Techniques and considerations," Journal of Medical Systems, vol. 44, no. 7, pp. 123, 2020.

[9] S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas, "Data preprocessing for supervised learning," International Journal of Computer Science, vol. 2, no. 2, pp. 111-117, 2020.

[10] L. Wang et al., "A hybrid CNN-LSTM model for cardiovascular disease prediction using symptom-based features," IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 6, pp. 1939-1950, 2021.

[11] A. Rajkomar et al., "Scalable and accurate deep learning with electronic health records," npj Digital Medicine, vol. 3, no. 1, pp. 1-10, 2020.