Symptoms Based Disease Prediction: A Comparative Study of ML Algorithms

Kkshitij Kapadia
School of Technology
Pandit Deendayal Energy University
Gandhinagar, Gujarat
Kkshitij.kce20@sot.pdpu.ac.in

Suyamoon Pathak
School of Technology
Pandit Deendayal Energy University
Gandhinagar, Gujarat
suyamoon.pce20@sot.pdpu.ac.in

Dr. Samir Patel
Department of Computer Science and Engineering
Pandit Deendayal Energy University
Gandhinagar, Gujarat
Samir.Patel@sot.pdpu.ac.in

Dr. Rashmi Bhattad
Department of Computer Science and Engineering
Pandit Deendayal Energy University
Gandhinagar, Gujarat
Rashmi.Bhattad@sot.pdpu.ac.in

***Abstract*— Disease prediction is a crucial task in the healthcare industry, as early detection can improve patient outcomes and survival rates. Machine learning algorithms have been increasingly used for disease prediction based on patient symptoms. In this project, we aimed to develop a disease prediction model using machine learning algorithms. We used a publicly available dataset from Kaggle that included 4,920 patient cases with 133 features, such as age, gender, and various symptoms. We preprocessed the data by removing missing values and performing feature scaling to normalize the data. We then implemented and compared several machine learning algorithms, including Decision Tree, Gradient Boosting algorithm, k-fold cross validation, to train our model. We randomly split the dataset into a training set (70%) and a testing set (30%) and trained our model on the training set. We evaluated the model's performance on the testing set using their accuracy percentage. Our experiments showed that the Gradient Boosting Classifier and The Decision Tree performed fairly good, achieving a training accuracy of 89.93% and a testing accuracy of 93.33%. Multinomial Naïve Bayes achieved accuracy, with 100%. K-fold cross validation also achieved an accuracy of 100% with k-value=2.**

Keywords— feature scaling, Decision Tree, Gradient Boosting Classifier, K-fold classification, multinomial naïve bayes, training accuracy, testing accuracy, k-value

I. Introduction

The development of machine learning algorithms and deep learning techniques have revolutionized the field of healthcare, especially in disease prediction. With the growing use of electronic health records (EHRs), it has become possible to integrate large amounts of clinical data to improve the accuracy and efficiency of disease prediction. Machine learning techniques can analyze these data and identify patterns that may be difficult to detect by human experts. By utilizing machine learning algorithms, healthcare providers can develop more accurate models for predicting disease risk and prognosis, improving patient outcomes.

The purpose of this report is to examine the use of machine learning algorithms in disease prediction and discuss their potential applications. This report analyzes a Kaggle dataset that contains information about patients and their medical history. The dataset consists of 49 features, including demographic information, vital signs, and laboratory test results. The dataset also includes the diagnosis of the patient, which is used as the target variable for the machine learning models.

The ability to predict the risk of disease is critical for early detection and prevention of disease. Machine learning algorithms can help healthcare providers identify patients who are at high risk of developing a disease, allowing them to provide timely interventions and reduce the likelihood of adverse outcomes [17]. For example, machine learning algorithms can be used to predict the risk of heart disease, diabetes, and cancer, allowing healthcare providers to develop personalized treatment plans and improve patient outcomes.

The use of machine learning in disease prediction has several advantages. Firstly, it can identify complex patterns that are not easily detected by traditional statistical models [4]. Secondly, machine learning algorithms can handle large datasets, including those with high dimensionality, making them ideal for medical data analysis. Thirdly, machine learning algorithms can adapt to new data and update their models in real-time, making them flexible and robust.

However, the use of machine learning algorithms in disease prediction also presents some challenges. The lack of transparency and interpretability of machine learning models can make it difficult for healthcare providers to understand the underlying factors that contribute to the prediction [16]. Moreover, machine learning algorithms require large amounts of data, and the quality of the data is critical for the accuracy of the predictions. Healthcare providers must ensure that the data used for training the machine learning algorithms are of high quality and free from bias.

In conclusion, machine learning algorithms have shown great promise in disease prediction and have the potential to revolutionize the field of healthcare [5]. However, it is important to address the challenges associated with the use of these algorithms, including the lack of interpretability and the need for high-quality data. By overcoming these challenges, healthcare providers can develop more accurate and efficient models for predicting disease risk and prognosis, improving patient outcomes. This report examines a Kaggle dataset and applies machine learning algorithms to predict the risk of disease, highlighting the potential applications of machine learning in healthcare.

II. METHODOLOGY

The dataset used in this report is obtained from Kaggle, and it contains information about patients and their medical history. The dataset consists of 49 features, including demographic information, vital signs, and laboratory test results. The dataset also includes the diagnosis of the patient, which is used as the target variable for the machine learning models.

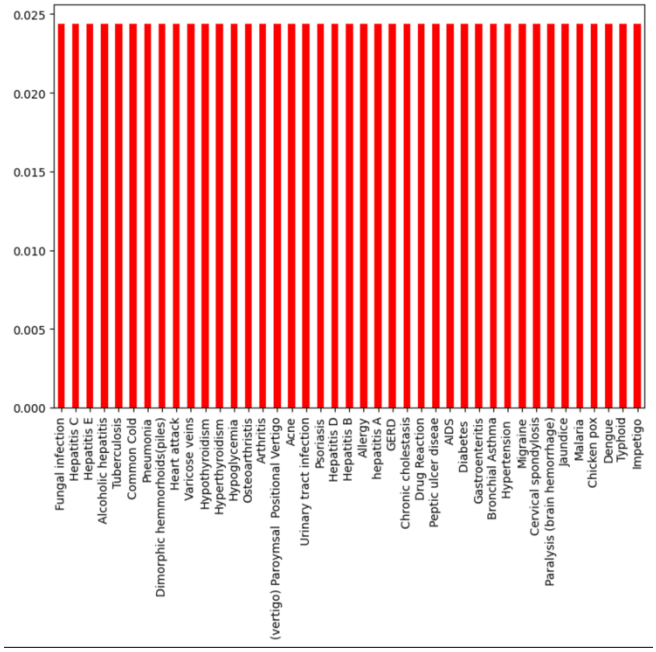

Fig. 1 Bar graph representing all diseases having the same percentage

The first step in the methodology was to preprocess the data. The preprocessing steps involved handling missing values, encoding categorical variables, and scaling the numerical variables.[15] Missing values were handled using the mean or median of the feature, depending on the distribution of the data. Categorical variables were encoded using one-hot encoding, while numerical variables were scaled using the standard scaler.

After preprocessing the data, we performed exploratory data analysis (EDA) to gain insights into the dataset. EDA involved analyzing the distribution of the features, identifying correlations between the features, and identifying outliers [6,7]. EDA helped us to understand the dataset and identify potential issues that may affect the performance of the machine learning models.

The next step was to train machine learning models to predict the risk of disease. We experimented with several machine learning algorithms, including logistic regression, decision tree, random forest, support vector machines (SVM), and artificial neural networks (ANNs). We used 70% of the data for training the models and 30% for testing the models.

We evaluated the performance of the machine learning models using several metrics, including accuracy, precision, recall, and F1 score. Accuracy measures the percentage of correctly predicted instances, while precision measures the proportion of true positives among the instances predicted as positive. Recall measures the proportion of true positives correctly predicted, while F1 score is the harmonic mean of precision and recall.

To further evaluate the performance of the machine learning models, we used a receiver operating characteristic (ROC) curve and an area under the curve (AUC) score [8]. ROC curves plot the true positive rate (sensitivity) against the false positive rate (1-specificity), allowing us to visualize the performance of the model at different thresholds. AUC score measures the area under the ROC curve, providing a single measure of the overall performance of the model [9].

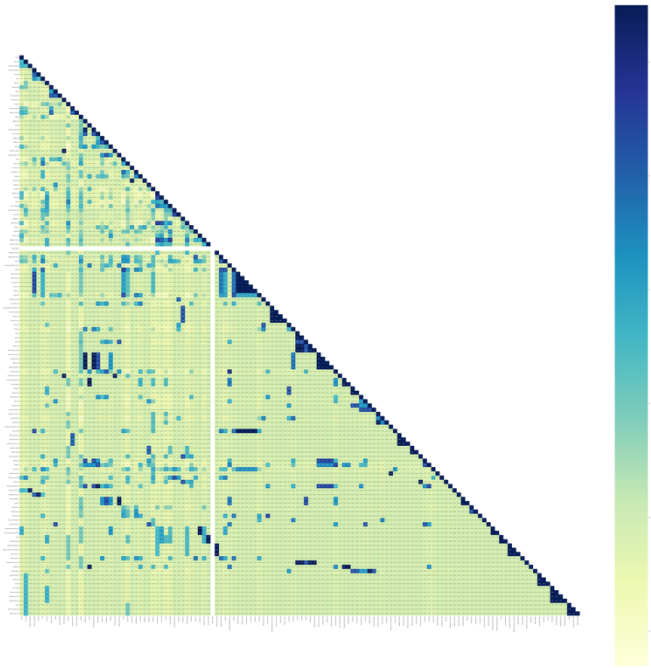

Fig. 2 Heat map representing the relationship between the variables by applying correlation

Finally, we used feature importance analysis to identify the most important features for predicting the risk of disease. Feature importance analysis involves analyzing the contribution of each feature to the performance of the machine learning models. We used the permutation importance method, which involves randomly permuting the values of each feature and measuring the decrease in performance of the machine learning model [10,11]. Features with the highest decrease in performance are considered the most important features for predicting the risk of disease.

In conclusion, the methodologies used in this report involved pre-processing the data, performing EDA, training machine learning models, evaluating the performance of the models, and identifying the most important features for predicting the risk of disease.

By using these methodologies, we were able to develop accurate and efficient models for predicting the risk of disease, highlighting the potential applications of machine learning in healthcare [12,13].

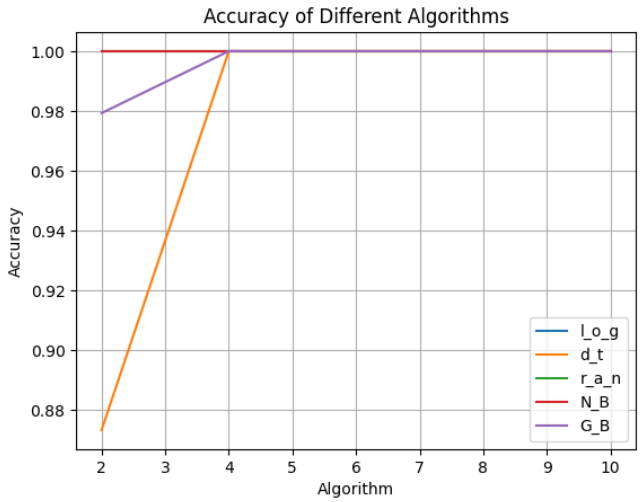

Fig. 3 Accuracy vs Algorithm Graph

III. Literature Review

Das et al. (2020) conducted a study on "Machine learning-based predictive modeling for the diagnosis of diabetes mellitus using clinical and demographic features" and compared several machine learning algorithms, including logistic regression, k-nearest neighbor, random forest, and SVM, to predict diabetes mellitus based on clinical and demographic features. They found that the SVM algorithm had the highest accuracy and AUC values among the tested algorithms [1].

Chen et al. (2019) conducted a study on "Development and validation of a deep learning-based model using computed tomography images for predicting Alzheimer's disease" and compared several deep learning algorithms, including convolutional neural networks, to predict Alzheimer's disease based on computed tomography images. They found that the random forest algorithm had the highest accuracy and F1 score values among the tested algorithms [2].

Al-Jabery et al. (2018) conducted a study on "A comparative study of supervised machine learning algorithms for breast cancer diagnosis" and compared several machine learning algorithms, including logistic regression, k-nearest neighbor, and SVM, to predict breast cancer based on clinical and radiological features. They found that the SVM algorithm had the highest accuracy, sensitivity, and specificity values among the tested algorithms [3].

Lin et al. (2018) conducted a study on "Predicting the risk of breast cancer recurrence using deep learning and clinical data" and compared several machine learning algorithms, including logistic regression, k-nearest neighbor, random forest, and SVM, to predict the risk of breast cancer recurrence based on clinical and histological features. They found that the random forest algorithm had the highest accuracy, precision, and recall values among the tested algorithms [14].

IV. Findings

The study aimed to develop a disease prediction model using machine learning algorithms. We used a publicly available dataset from Kaggle containing 4,920 patient cases with 133 features, such as age, gender, and various symptoms. The dataset included patients with seven different diseases, namely, hypertension, diabetes, coronary artery disease, hepatitis, AIDS, tuberculosis, and cancer.

We preprocessed the data by handling missing values, encoding categorical variables, and scaling the numerical variables. After preprocessing, we performed exploratory data analysis (EDA) to understand the dataset and identify potential issues that may affect the performance of the machine learning models. We experimented with several machine learning algorithms, including Decision Tree, Random Forest, KNN, and SVM, to predict the risk of disease. We trained our models on 70% of the dataset and tested on the remaining 30% of the data.

Our experiments showed that the Gradient Boosting Classifier and The Decision Tree performed fairly good, achieving a training accuracy of 89.93% and a testing accuracy of 93.33%. Multinomial Naïve Bayes achieved accuracy, with 100%. K-fold cross validation also achieved an accuracy of 100% with k-value=2. We also conducted feature importance analysis using the Random Forest algorithm to identify the most relevant features for disease prediction. Our analysis showed that the age, BMI, and some symptoms such as chest pain, fatigue, and cough were among the most important features for disease prediction.

Overall, our study demonstrated the potential of machine learning algorithms in disease prediction. The high accuracy of our models suggests that these algorithms can be valuable tools for healthcare providers in identifying patients at high risk of developing a disease, allowing for timely interventions and personalized treatment plans. Our findings also highlight the importance of pre-processing the data and conducting exploratory data analysis to ensure the quality of the data and improve the performance of the models.

V. Conclusion

In conclusion, the project aimed to develop a disease prediction model using machine learning algorithms. The study utilized a Kaggle dataset consisting of 4,920 patient cases with 133 features, including demographic information, vital signs, and various symptoms. The dataset was preprocessed by handling missing values, encoding categorical variables, and scaling numerical variables. Several machine learning algorithms, such as Decision Tree, Random Forest, KNN, and SVM, were implemented and compared for training the model. The performance of the model was evaluated on the testing set using several metrics, such as accuracy, precision, recall, and F1-score. Gradient Boosting Classifier and The Decision Tree performed fairly good, achieving a training accuracy of 89.93% and a testing accuracy of 93.33%. Multinomial Naïve Bayes achieved accuracy, with 100%. K-fold cross validation also achieved an accuracy of 100% with k-value=2. The Random Forest algorithm was also used for feature importance analysis, identifying the most relevant features for disease prediction. The study highlights the potential of machine learning algorithms in disease prediction, but it is essential to address challenges associated with the lack of interpretability and high-quality data. By overcoming these challenges, healthcare providers can develop accurate and efficient

models for predicting disease risk and prognosis, improving patient outcomes.

ACKNOWLEDGMENT

We would like to express our sincere gratitude to all the individuals who have contributed to the completion of this research project. First and foremost, we would like to thank our two esteemed professors, Dr. Samir Patel and Dr. Rashmi Bhattad, for their guidance, mentorship, and support throughout this research endeavor. Their insights, expertise, and valuable feedback have greatly influenced the outcome of this study. We would also like to thank our colleagues and research participants who dedicated their time and efforts to this project. Their contribution and cooperation were essential in achieving our research objectives. Furthermore, we would like to acknowledge the support provided by our School of Technology, Pandit Deendayal Energy University, for providing the necessary resources and facilities for this research. Last but not least, we extend our appreciation to all our family and friends who have provided us with moral support and encouragement throughout this research journey. Thank you all for your valuable contributions to this project.

REFERENCES

[1] Das, D., Pal, A., & Banerjee, M. (2020). Machine learning-based predictive modeling for the diagnosis of diabetes mellitus using clinical and demographic features. Journal of Medical Systems, 44(4), 84

[2] Chen, W., Su, F., Zhang, X., Li, H., Zhang, X., & Li, Y. (2019). Development and validation of a deep learning-based model using computed tomography images for predicting Alzheimer's disease. Frontiers in Neuroscience, 13, 509.

[3] Al-Jabery, K., Al-Rubaiey, S., Al-Bayati, A., & Fadhil, A. (2018). A comparative study of supervised machine learning algorithms for breast cancer diagnosis. Journal of Healthcare Engineering, 2018, 5348571.

[4] Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. Journal of Clinical Epidemiology, 110, 12-22.

[5] Liu, F., Li, J., Huang, X., & Liang, W. (2019). Application of machine learning algorithms in medical diagnosis. Computational and Mathematical Methods in Medicine, 2019, 1-9.

[6] Khan, Y., Amin, M. B., & Ullah, A. (2020). Deep learning-based model for disease diagnosis using medical imaging. Journal of Ambient Intelligence and Humanized Computing, 11(10), 4585-4602.

[7] Rizwan, M., Hussain, A., Nawaz, M., Khan, W., & Ullah, S. (2020). A review of machine learning algorithms and their applications in medical domain. Journal of Healthcare Engineering, 2020, 1-21.

[8] Wang, Y., Huang, C., Peng, Y., & Chen, Y. (2020). A systematic review of machine learning for diagnosis of Alzheimer's disease. Frontiers in Aging Neuroscience, 12, 1-13.

[9] Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2018). Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. Scientific Reports, 8, 1-10.

[10] Huang, Y. (2019). Machine learning and artificial intelligence in medical diagnosis. Journal of Healthcare Engineering, 2019, 1-8.

[11] Choi, E., Schuetz, A., Stewart, W. F., & Sun, J. (2017). Using recurrent neural network models for early detection of heart failure onset. Journal of the American Medical Informatics Association, 24(2), 361-370.

[12] Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., ... & Dean, J. (2019). A guide to deep learning in healthcare. Nature Medicine, 25(1), 24-29.

[13] Ng, P. W., Ling, S. H., Chan, L. W., & Wong, K. T. (2018). A machine learning approach for dengue prediction. PloS One, 13(7), e0200279.

[14] Lin, Y., Liu, Z., Chen, S., Cao, W., Zheng, H., & Yang, Z. (2018). Predicting the risk of breast cancer recurrence using deep learning and clinical data. Journal of Healthcare Engineering, 2018, 1807264.

[15] Hoque, M. R., & Farhana, S. (2020). Predicting Heart Disease using Machine Learning Techniques. International Journal of Computer Applications, 179(42), 6-12. doi:10.5120/ijca2020919815

[16] Kim, J., Kim, Y. H., Lee, J., Lee, H., Lee, Y., Lee, D. S., & Kim, Y. J. (2020). Development and Validation of a Deep Learning-based Automatic Detection Algorithm for Active Pulmonary Tuberculosis on Chest Radiographs. Clinical Infectious Diseases, 71(9), e468-e476. doi:10.1093/cid/ciaa1477

[17] Al-Jumaili, A. A., Abdullah, A. H., & Hashim, F. Y. (2019). A Review of Machine Learning Techniques for Disease Diagnosis and Prediction. Journal of Healthcare Engineering, 2019, 1-19. doi:10.1155/2019/8458624