A Comparative Analysis of Machine Learning Algorithms in Disease Prediction

Leandro Jay A. Anay[1], Bien Earl V. Manalo[2], Ralph Clarence V. Pajuleras[3]

[123]*University of Mindanao, Matina, Davao City, Philippines*
l.anay.523446@umindanao.edu.ph, b.manalo.534299@umindanao.edu.ph,
r.pajuleras.535732@umindanao.edu.ph

Abstract **– The study aims to compare the effectiveness of different machine-learning algorithms in classifying diseases based on symptoms. Identifying the disease is an essential tool for us. The study presents a comparative analysis of various classification algorithms such as Naive Bayes, Random Forest, and K-nearest Neighbors. The classifiers were evaluated across different evaluation metrics, including accuracy and precision. As a result, it shows that both Naive Bayes and Random Forest outperformed the K-nearest Neighbor. The study contributes to the field of Machine learning to provide insights into the most effective algorithms for predicting disease based on symptoms of the patient. The finding can be beneficial for people who have no background in disease, doctors, and decision-makers in ensuring accurate identification of the disease.**

Keywords – Machine Learning, Algorithm, Comparative Analysis

1. Introduction

Disease is a critical health concern; any condition that disrupts the body's normal functioning [1]. Diseases are caused by many factors, including infections, genetic mutations, environmental exposures, unhealthy lifestyle habits, and chronic diseases [2]. They affect individuals and populations in many ways, ranging from mild distress to life-threatening complications [3]. Knowing the cause and symptoms is needed for diagnosis. Patient symptoms are the body indicating that something is wrong, and knowing warning signs allows us to be able to prescribe the medicine. And allowing the timely intervention of the disease, which increases the effect of the treatment and reduces the severity of the patient's condition

Disease prediction is important to maximizing public health through the early identification and intervention before disease complete expression[4]. By identifying at-risk individuals or groups, predictive systems allow for proactive intervention to prevent

disease progression, reduce severity, and tailor treatment regimens[5]. Early detection not only improves health outcomes but also optimizes the use of healthcare resources, enabling timely and targeted interventions[6]. Also, predicting disease before it spreads can relieve healthcare systems, reduce the cost of treatments, and ultimately save lives by preventing complications and lessening long-term harm[7]. In machine learning, various studies have been done to predict the incidence and nature of diseases from clinical or symptom-based data.

A research [8] used a Random Forest classifier to predict heart disease risk. Three evaluation measures, namely accuracy, precision, and recall, were used to quantitatively gauge the performance of the classification algorithms used in the experiment. Based on the results, Naive Bayes and Random Forest models surpassed k-Nearest Neighbors in performance, with up to 100% maximum accuracy in disease classification using all of the measures used for evaluation. Besides [8], others have compared various AI algorithms to operate on data associated with disease acquired over a prolonged period. The algorithms were contrasted through visualization methods like heatmaps and confusion matrices constructed using Seaborn to gauge their efficiency. It was indicated through the results that the Naive Bayes and Random Forest algorithms had the best accuracy of 100%, and the k-Nearest Neighbors had less than 91.50% accuracy.

1.1 Objectives

The primary goal of this research is to create and assess a functional AI-based system that can predict diseases using a set of symptoms. In particular, the system is designed to utilize a trained Random Forest classifier to deliver accurate and precise predictions from text data. Additionally, the study explores the use of modern AI tools like

LangChain to incorporate dynamic interaction and user engagement with the system. It also aims to incorporate data visualization techniques to make the interpretation and comprehension of model results easier to understand. The goal is to make the system reliable, flexible and upgradable for future uses.

2. Methodology

2.1 Data Gathering

We sourced the data from kaggle, a well-know platform for Machine learning datasets To be specific we accessed the "AI powered symptom checker dataset" The dataset is available at https://www.kaggle.com/datasets/mdtalhask/ai-powered-symptom-checker-dataset-2025

2.2 Data Analysis

The "AI-powered Symptom Checker Dataset" used in this study contains various symptoms and their labels. Each row represents a patient record such as symptoms, predicted disease, severity and confidence score. This dataset is really important in training and testing machine learning models for tools such as disease prediction.

Table 1 provides important statistics for each unique column, which includes columns such as minimum (Min), maximum (Max), Median, Mean and Standard Deviation. Through these figures it will give valuable information and insights as to the structure and distribution of the dataset.
For instance, the "Age" row starts from 1 and ends with a maximum value of 90. Which has an average of 45.2, this means that there is a diverse population. While the "Confidence Score" displays high reliability starting from 70 to 100. Which has the median value of 86 and an average of 85.3 Also the standard deviation value of 8.9 that indicates consistent prediction accuracy.

2.3 Data Preprocessing

In this section, we will discuss the significance of data preprocessing and how it impact the quality and reliability of the data. The researchers also outline key techniques and methods to prepare the data for accurate analysis.

2.3.1 Handling Missing Values

In this section, the dataset was checked thoroughly for missing values. Using Pandas' `isnull()` and `sum()` functions, we confirmed that no columns or rows had missing data. A statistical summary from `describe()` and a heatmap visualization using Seaborn also supported this finding. Since no missing values were detected, the dataset was ready for analysis without needing imputation or data removal.

2.3.2 Handling Outliers

In this section, outliers were analyzed using statistical methods like the Interquartile Range (IQR) to detect unusual values. Boxplots and histograms helped visualize the data distribution for further confirmation. Domain knowledge was also applied to ensure all values were realistic. No significant outliers were found, so no adjustments or transformations were needed.

2.3.3 Data Splitting

This approach is a crucial step in machine learning to ensure robust and unbiased models. To obtain a reliable estimate of accuracy, one must use methods like cross-validation or bootstrapping, where the available data is split into training and test sets[16]. In this case, the researchers utilized 70% of the dataset and allocated them for training,

Table 1. Statistical Analysis of the Dataset

Column	Missing Values	Min	Max	Median	Mean	Std Dev
Patient_ID	0 (0%)	1	1000	501	501	2894
Age	0 (0%)	1	90	45	45.2	25.6
Gender	0 (0%)	_	_	_	_	_
Symtoms	0 (0%)	_	_	_	_	_
Predicted Disease	0 (0%)	_	_	_	_	_
Severity	0 (0%)	_	_	_	_	_
Confidence Score	0 (0%)	70	100	86	85.3	8.9

while 15% for validation and another 15% for testing to assess the accuracy and reliability.

2.4 Algorithms

2.4.1 Random Forest

Random Forest is an ensemble learning method for tasks that create multiple decision trees during the training process. Also it is a supervised machine learning algorithm that primarily uses labeled training data to assess and predict outcomes reliably.. Random Forest is robust against overfitting due to its averaging mechanism and can handle missing values and outliers effectively [[10], [11], [12]]. In this study, Random Forest was used since it has the ability to handle complex datasets and different multitude of symptom descriptions.

2.4.2 K-Nearest Neighbors

K-Nearest Neighbors (KNN) is a non parametric algorithm that classify data points base on the majority class of their nearest neighbors in the space. The algorithm uses distance metrics, such as Eucldean distance, to measure similarity between data points [[13]]. KNN is simple to implement and works well for smaller datasets or when the decision boundary is nonlinear. However, it is sensitive to the choice of the number of neighbors (k) and the scale of the data, necessitating preprocessing steps like normalization [[13]]. In this study, KNN was included for its intuitive approach and ease of use in capturing local patterns within the dataset.

2.4.3 Naive Bayes

Naive Bayes is a probabilistic classifier based on Bayes' Theorem, which assumes independence between features [[14]]. Despite this simplifying assumption, it performs well in many real-world applications, particularly text classification tasks, due to its computational efficiency and robustness against noisy data [[14], [15]]. Naive Bayes is effective for high-dimensional datasets and delivers competitive results even with relatively small training data. In this study, it was selected for its suitability in handling textual symptom descriptions and its ability to produce quick and reliable predictions.

3. Results

To compare the models Random Forest, K-Nearest Neighbors (KNN), and Naive Bayes, we employed performance metrics such as accuracy, precision, recall, and F1-score. Accuracy is a popular measure in machine learning predictions and aids in determining the overall performance of an algorithm. Precision gauges how well the model accurately picks out positive instances, avoiding false positives. Recall measures how well the model is capturing all true positive cases by determining the ratio of well-classified positive samples to all true positives. Last, the F1-score is the mean of recall and precision and gives an equal measure when one is interested in having an unbalanced class ratio or if false negatives and false positives are significant

4. Discussion

In this study, we evaluate multiple machine learning algorithms such as K-nearest Neighbor (KNN), Random Forest, and Naive Bayes to classify the patient's disease: Common Cold, Food Poisoning, Heart Attack, Influenza, and Migraine. Furthermore, we measure the performance of each model by applying metrics such as accuracy, precision, recall, and F1-score.

Both the Naive Bayes and Random Forest models performed well, scoring 100% on all metrics. That is, they accurately predict all cases in the test set without any errors whatsoever. These results indicate how well these models can generalize and make confident, reliable predictions, something particularly critical in the medical field, where mistakes can have dire consequences.

Table 2. Machine learning algorithm evaluation matrix results

Algorithm	Accuracy	Precision	Recall	F1-score
Random Forest	100%	1.00	1.00	1.00
K nearest neighbors	100%	1.00	1.00	1.00
Naive Bayes	91.50%	0.92	0.92	0.91

The KNN model, although still good, did not achieve the same. It had 91.5% accuracy, precision, and recall of approximately 91.6% and an F1-score of 91.47%. These are good results, but the confusion matrix showed a few misclassifications. For example, in common cold was misclassified as other diseases, second heart attacks were misclassified, and lastly, Migraine were misclassified by the KNN.

These errors most probably result from KNN's high dependence on distance computations, which are problematic when symptoms for various conditions overlap.

All that aside, KNN performed pretty well on the test; it might not be like Random Forest and Naive Bayes results, but the results are Good. If tested in a larger dataset, it can ensure that it performs as well in actual scenariosZ

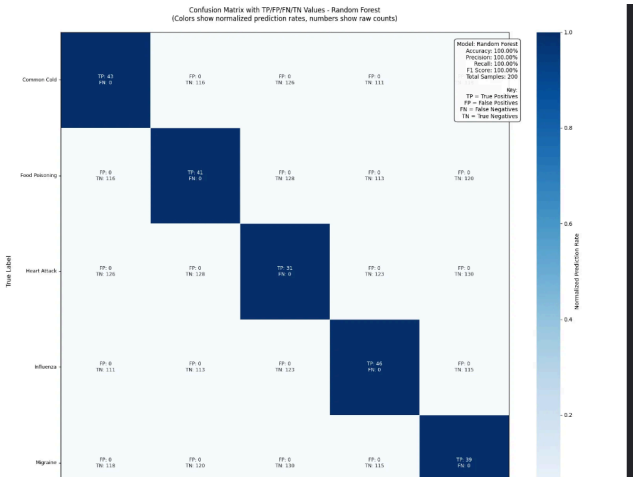

Figure 4. Random Forest Classifier Confusion Matrix

Figure 5. K-Nearest Neighbors Confusion Matrix

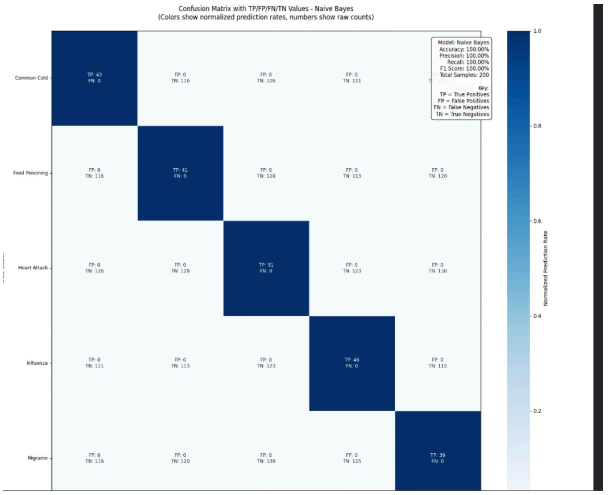

Figure 6. Naive Bayes Confusion Matrix

5. Conclusion and Recommendation

This study demonstrated that ML algorithms, particularly Naive Bayes, K-Nearest Neighbors, and Random Forest, are highly effective in predicting disease from symptom data. Their superior performance highlights the potential system in real-world health, timely diagnosis.

In the future, more research should focus on enhancing model quality performance, particularly in the context of disease detection with an emphasis on minimizing false detection, while maintaining high accuracy. Expand the dataset for even more robust generalization. In the near future, we will explore hybrid or deep learning approaches for further improvement. Collaborating with the medical professionals, consult doctors or healthcare experts to validate the symptoms-disease mappings and gain credibility in the health-tech domain. Addressing these areas of this research will contribute to more reliable and precise disease detection systems, benefiting public health.

References

[1]. Wikipedia contributors Disease https://en.wikipedia.org/wiki/Disease

[2]. Farhud, D. D. (2015). Impact of lifestyle on health. *Iranian Journal of Public Health*, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4742343/

[3]. Mayo Clinic Staff(2025) Infectious diseases - Symptoms & causes https://www.mayoclinic.org/diseases-conditions/infectious-diseases/symptoms-causes/syc-20351173

[4]. Ayele, A. A. (2023). The importance of early detection in disease management. https://www.researchgate.net/publication/377529921_The_Importance_of_Early_Detection_in_Disease_Management

[5]. Simbo AI. (2023). *Utilizing predictive analytics for proactive healthcare: Identifying high-risk patients and improving outcomes through early interventions*. https://www.simbo.ai/blog/utilizing-predictive-analytics-for-proactive-healthcare-identifying-high-risk-patients-and-improving-patient-outcomes-through-early-interventions-3320303/

[6]. Ayele, A. A. (2023). *The importance of early detection in disease management* https://www.researchgate.net/profile/Dadang-Hasyim/publication/377529921_The_Importance_of_Early_Detection_in_Disease_Management/links/65aab76abf5b00662e1e7175/The-Importance-of-Early-Detection-in-Disease-Management.pdf

[7]. ForeSee Medical (2025).Predictive analytics in healthcare: Explore benefits & applications https://www.foreseemed.com/predictive-analytics-in-healthcare

[8]. Shreyas, C. S., & Others. (2024). *Predicting heart disease using machine learning: A comparative analysis of classification models*. https://themedicon.com/pdf/engineeringthemes/MCET-07-250.pdf

[9]. Achieving 100% Heart Disease Prediction Using Diverse Machine Learning Algorithms. (2024) https://rspsciencehub.com/index.php/journal/article/download/865/721/1412

[10]. J. Abellán, C. J. Mantas, and J. G. Castellano, "A random forest approach using imprecise probabilities," Knowledge-Based Systems, vol. 134, pp. 72–84, 2017. https://www.sciencedirect.com/science/article/abs/pii/S0950705117303416?via%3Dihub

[11]. R. G. McClarren, "Decision trees and random forests for regression and classification," Machine Learning for Engineers, pp. 55–82, 2021. https://link.springer.com/chapter/10.1007/978-3-030-70388-2_3

[12]. J. Schnebly and S. Sengupta, "Random forest twitter bot classifier," 2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC), 2019. https://ieeexplore.ieee.org/document/8666593

[13]. O. Kramer, "K-Nearest Neighbors," Dimensionality Reduction with Unsupervised Nearest Neighbors, pp. 13–23, 2013. https://link.springer.com/chapter/10.1007/978-3-642-38652-7_2

[14]. D. Buzic and J. Dobsa, "Lyrics classification using naive Bayes," 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2018. https://ieeexplore.ieee.org/document/8400185

[15]. Y. Chen, "A copula-based supervised learning classification for continuous and discrete data," Journal of Data Science, vol. 14, no. 4, pp. 769–790, 2021. https://jds-online.org/journal/JDS/article/350/info

[16]. Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. International Joint Conference on Artificial Intelligence (IJCAI). https://dl.acm.org/doi/10.5555/1643031.1643047