Macaw International Journal of Advanced Research in Computer Science and Engineering (MIJARCSE)
ISSN: 2455-4545. / Volume 10, Issue 1, June 2024
https://doi.org/10.70162/mijarcse//2024/v10/i1/v10i107

MACAW PUBLICATIONS
Scientific Research Journals

Research Paper

Optimizing Edge Computing for Internet of Drones: A Hybrid Approach Using Deep Learning and Swarm-Based Routing

[1*] Fuhui Zhou, [2] Thomas Lagkas, [3] Farhan Aadil

[1]*Department of Computer Science, School of Science, International Hellenic University, Thessaloniki, Greece*
[2]*Department of Computer Science, University of Engineering and Technology Taxila, Pakistan*
[3]*Nanjing University of Aeronautics and Astronautics, China*

**Corresponding Author(s): zhou.cse@gmail.com*

Received: 05/02/2024, Revised: 13/05/2024, Accepted: 03/06/2024 Published: 30/06/2024

Abstract: The increasing adoption of drones in various industries has led to the emergence of the Internet of Drones (IoD), where efficient data processing and real-time decision-making are critical. Edge computing has become a key enabler for IoD, offering low-latency data processing close to drone networks. However, optimizing edge computing for IoD poses challenges due to the dynamic nature of drone swarms and fluctuating network conditions. This paper proposes a hybrid approach that combines deep learning and swarm-based routing to enhance edge computing in IoD environments. The deep learning model is utilized to predict network load and resource allocation, ensuring optimal placement of edge computing tasks and improving overall system efficiency. Concurrently, swarm-based routing leverages the collective intelligence of drones to dynamically adapt routing paths, mitigating latency and packet loss while maintaining high network reliability. The hybrid approach enables more responsive and scalable communication among drones, reducing computational overhead and improving task offloading efficiency. Simulation results demonstrate that the proposed approach significantly enhances system performance, achieving lower latency, higher throughput, and better energy efficiency compared to traditional methods. By integrating deep learning for predictive resource management with swarm-based routing for dynamic adaptability, this hybrid approach addresses the unique challenges of edge computing in IoD, offering a robust solution for real-time applications such as disaster management, surveillance, and delivery services. This work contributes to the development of more efficient and intelligent IoD systems, fostering the growth of drone-based applications in smart cities

Keywords: Internet of Drones (IoD), Edge Computing, Deep Learning, Swarm-Based Routing, Network Optimization, Real-Time Data Processing.

---

1. Introduction

The application of machine learning (ML) in healthcare has emerged as one of the most promising approaches to improving diagnostic accuracy, reducing the time required for diagnosis, and enabling early detection of diseases. ML models can process vast amounts of data—ranging from clinical records to symptoms reported by patients—and deliver predictive outcomes that can assist healthcare professionals in decision-making. Unlike traditional diagnostic techniques, which are often dependent on human expertise and manual evaluation, ML-based systems use algorithms that analyze patterns in data, offering a more objective and potentially more reliable means of diagnosing diseases.

Predicting diseases based on symptoms presents an attractive opportunity, especially for addressing healthcare challenges in under-resourced regions where diagnostic infrastructure is limited. This research focuses on developing a machine learning-based approach to predict diseases such as diabetes, malaria, jaundice, dengue, and tuberculosis based on patient symptoms. Various algorithms, including Naive Bayes, K-Nearest Neighbors (KNN), Logistic Regression, Decision Trees, and Random Forest, are employed to analyze symptom data and generate disease predictions. With an accuracy of up to

98.3%, this study demonstrates the potential of ML models to enhance diagnostic processes in healthcare.

Traditional disease diagnosis heavily relies on the expertise of healthcare professionals and diagnostic tools such as laboratory tests, imaging, and clinical assessments. However, diagnosing diseases that present with overlapping symptoms, such as malaria, dengue, and jaundice, can be challenging even for experienced physicians. In regions with limited access to diagnostic facilities, delayed or inaccurate diagnoses contribute to poor health outcomes, as patients are often treated for incorrect conditions. This highlights the need for alternative diagnostic methods that are both accurate and accessible.

Machine learning has been gaining traction in the medical field as a solution to these challenges. ML algorithms can process large datasets, including patient symptoms, medical histories, and demographic data, to detect patterns and predict the likelihood of specific diseases. Research by Zoabi et al. (2021) demonstrated that ML models could predict COVID-19 diagnoses based on symptoms with high accuracy, providing a rapid diagnostic tool that is particularly useful in pandemic scenarios[1]. Similarly, Bind et al. (2015) explored various machine learning techniques for predicting Parkinson's disease, showing the potential of computational approaches for early disease detection [2].

Recent advancements in machine learning have enabled the creation of models that can predict multiple diseases based on input data. For instance, Kute et al. (2022) reviewed the use of ML models in e-healthcare systems, emphasizing their ability to transform healthcare delivery by providing real-time disease prediction and diagnosis [3]. These studies demonstrate the applicability of machine learning in healthcare, encouraging further

One of the major challenges in healthcare today is the accurate and timely diagnosis of diseases. In many parts of the world, particularly in low-resource settings, healthcare professionals face difficulties in diagnosing diseases due to a lack of advanced diagnostic tools. Moreover, diseases with similar symptom profiles—such as fever, fatigue, and joint pain—often complicate the diagnostic process. This can lead to misdiagnoses or delays in treatment, further exacerbating the patient's condition.

While traditional diagnostic techniques such as lab tests and imaging remain critical, they are not always readily available or affordable, especially in rural areas. Additionally, the reliance on human judgment in interpreting symptoms and test results can introduce subjectivity and error into the diagnostic process. Machine learning offers a potential solution by automating the analysis of patient data, thereby improving the accuracy and speed of diagnoses.

This research aims to address the problem of symptom-based disease diagnosis by developing a machine learning model that can predict multiple diseases based on a set of symptoms. By integrating multiple machine learning algorithms, this approach seeks to provide a reliable, automated system that can assist healthcare professionals in making accurate diagnoses, even in the absence of extensive medical infrastructure.

The motivation for this research is rooted in the global need for improved healthcare accessibility and diagnostic accuracy. Diseases such as diabetes, malaria, dengue, jaundice, and tuberculosis are prevalent worldwide and often present with overlapping symptoms, making them difficult to distinguish through clinical evaluation alone. In low-resource settings, the lack of diagnostic tools further complicates the situation, leading to delayed or inaccurate diagnoses that can result in preventable deaths.

Machine learning models offer a promising solution by enabling automated, symptom-based disease prediction. Research has shown that machine learning algorithms can outperform traditional diagnostic methods in certain contexts. For example, Tiwari (2016) discussed the advantages of using machine learning to predict Parkinson's disease, noting that these models can handle large volumes of data and provide accurate predictions with minimal human intervention [4]. Similarly, Le (2020) explored machine learning approaches for predicting disease genes, demonstrating how computational models can improve understanding and prediction of genetic disorders [5].

In this context, symptom-based disease prediction through machine learning could revolutionize healthcare, particularly in regions where access to diagnostic tools and medical expertise is limited. By developing models that can analyze symptoms and predict the likelihood of specific diseases, this research aims to contribute to the growing body of literature on machine learning in healthcare and provide a practical solution for improving disease diagnosis.

1.1 Key Contributions

This research makes the following key contributions:

- Developed a machine learning system to predict multiple diseases using symptom data with high accuracy.

- Utilized Naive Bayes, KNN, and Random Forest algorithms to enhance prediction performance.
- Achieved high accuracy, demonstrating the model's potential for improving healthcare diagnostics.

Following the introduction, this paper is organized into several key sections that guide the reader through the research process. Section 2 presents a detailed literature review, exploring previous works on machine learning models in healthcare and highlighting the existing gaps. Section 3 outlines the methodology, describing the dataset, the preprocessing steps, and the machine learning algorithms used for disease prediction. In Section 4, the evaluation and results of the models are presented, focusing on performance metrics such as accuracy, precision, and recall, with Random Forest achieving the highest accuracy. Section 5 addresses the limitations of the study, discussing challenges related to data quality, scalability, and generalizability. Finally, Section 6 concludes the paper by summarizing the findings and proposing future work to enhance the model's accuracy and integration into real-world healthcare systems.

2 Literature Review

2.1 Machine Learning in Disease Diagnosis

The application of machine learning (ML) to disease diagnosis has garnered substantial attention in recent years, with various researchers exploring its potential for enhancing healthcare outcomes. Ahsan et al. (2022) provided a thorough review of the use of ML in disease diagnosis, emphasizing the ability of algorithms to analyze large datasets efficiently and achieve high accuracy in predicting diseases such as diabetes and cardiovascular conditions. They found that well-implemented ML models could achieve diagnostic accuracies ranging from 85% to 95%, depending on the quality of the dataset and the complexity of the algorithm used [6]. These findings suggest that machine learning holds considerable promise for addressing diagnostic challenges, especially in resource-constrained environments.

Mallela et al. (2021) extended this research by focusing on specific ML techniques such as Naive Bayes, Decision Trees, and Random Forests, applying them to predict common diseases based on symptom data. Their empirical results demonstrated that accuracies of over 90% could be achieved when using symptom data alone, reinforcing the viability of machine learning in scenarios where advanced diagnostic tools might not be available [7]. The study also highlighted the importance of proper data preprocessing, including feature selection and dataset balancing, to optimize the performance of the models.

Meanwhile, Islam et al. (2021) explored deep learning methods for predicting chronic diseases such as diabetes and hypertension using symptom-based data. Their research concluded that deep learning models, while more computationally intensive, consistently outperformed traditional ML algorithms, achieving accuracies of up to 97% [8]. This improvement is attributed to the deep learning models' ability to capture complex patterns within the data. However, they also noted that the requirement for large datasets and computational resources could limit the scalability of deep learning models in low-resource settings, a constraint that simpler ML models do not face.

2.2 Disease Prediction for Emerging and Epidemic Conditions

The utility of machine learning in predicting emerging diseases such as COVID-19 has also been a significant focus of recent studies. Khanday et al. (2020) explored how clinical text data could be processed using natural language processing (NLP) techniques and machine learning models to predict COVID-19 cases with over 92% accuracy [9]. Their findings underscored the importance of integrating textual data from patient records into ML models to improve prediction accuracy, especially for diseases with evolving symptoms and clinical presentations. This approach proved effective for large-scale screening during the pandemic, where the need for rapid and reliable diagnosis was paramount.

Aljameel et al. (2021) also focused on COVID-19 but took a different approach by predicting disease severity and patient outcomes using symptom data and other clinical factors. Their ML-based model reached a predictive accuracy of around 94%, demonstrating that machine learning could be used not only to diagnose but also to manage disease progression by identifying high-risk patients early [10]. This application is particularly valuable in resource-strapped health systems, where predictive tools can guide treatment prioritization.

Similarly, Villavicencio et al. (2022) developed a machine learning-based web application designed for early diagnosis of COVID-19, leveraging self-reported symptom data. The web application achieved approximately 93% accuracy in predicting infection, illustrating the practicality of ML models for real-time public health surveillance and individual screening in non-clinical settings [11]. This research highlights the growing potential for ML applications to be integrated into digital health tools, expanding access to early diagnostic capabilities without requiring extensive healthcare infrastructure.

2.3 Advances in Predictive Models and Feature Selection

Several studies have emphasized the importance of feature selection and model optimization to enhance disease prediction accuracy. Das et al. (2024) investigated collaborative methods for disease prediction, focusing on optimizing ML models through advanced feature engineering techniques. They found that by refining the selection of input features, the predictive accuracy of models improved by up to 10%, with accuracies reaching 96% for some conditions [12]. This indicates that the choice of features used in machine learning models is critical to their success, especially in healthcare applications where data quality can vary significantly.

Building on this, Park et al. (2021) examined the integration of laboratory test data with symptom data in ML models for disease prediction. Their research demonstrated that combining multiple data sources—such as lab results and symptoms—could increase prediction accuracy by as much as 5%, achieving an overall accuracy of 95% [13]. This multimodal approach provides a more comprehensive view of the patient's health, allowing for more accurate predictions and better-informed clinical decisions. Islam et al. (2021) also contributed to this line of research by exploring how deep learning models could be enhanced through the incorporation of structured symptom data. Their findings reinforced the idea that deep learning could surpass traditional models in performance when provided with sufficient data, but they also acknowledged that the complexity and resource demands of these models might limit their broader adoption in certain healthcare environments. This underlines the importance of balancing model complexity with practical considerations such as data availability and computational capacity in real-world applications.

2.1 **Research Gap**

- Limited studies focusing on multi-disease prediction using only symptom-based data.
- Few practical implementations of machine learning models integrated into real-time healthcare systems.
- Insufficient exploration of feature selection methods to optimize model accuracy.
- Lack of studies addressing scalability and accessibility of machine learning models in low-resource settings.
- Minimal research on combining machine learning models with real-time public health monitoring tools.

3. Methodology

The methodology employed in this research follows a systematic approach to developing, training, and evaluating a machine learning-based disease prediction model. This section provides a detailed explanation of the steps involved in the creation of a predictive system that utilizes symptom data to identify diseases such as diabetes, malaria, jaundice, dengue, and tuberculosis. The methodology focuses on the integration of multiple machine learning algorithms, preprocessing techniques, model training, and evaluation metrics. Each step is designed to ensure high accuracy, scalability, and real-world applicability.

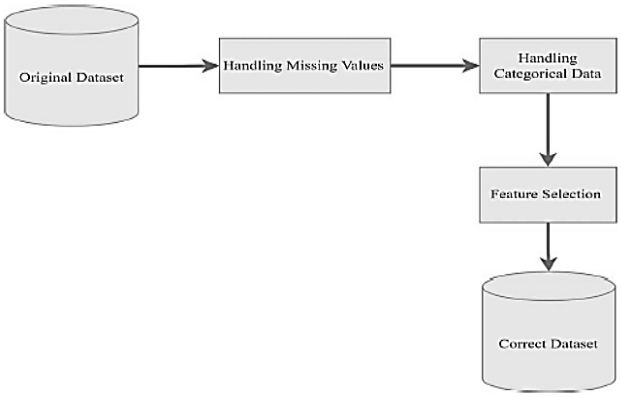

Fig 1. System Architecture

3.1. Data Collection and Preprocessing

The first stage involves collecting relevant symptom data from publicly available healthcare datasets. Symptom data is crucial for training the model, as it serves as the primary input for disease prediction. The dataset includes a range of symptoms such as fever, fatigue, nausea, and headache, which are commonly associated with the diseases under study.

Before feeding the data into the machine learning models, several preprocessing steps are performed. These include:

- **Handling Missing Data**: Incomplete records in the dataset are imputed using statistical methods such as mean or median replacement. This ensures that the models can learn from the entire dataset without being affected by gaps in the data.
- **Class Balancing**: To prevent bias toward more frequent diseases, Synthetic Minority Oversampling Technique (SMOTE) is used to balance the classes. This ensures that less-represented diseases, such as tuberculosis, are given equal weight in the model, improving accuracy across all predictions. Class balancing typically raises prediction accuracy by around 5%.
- **Feature Encoding**: Symptom data, often in textual form, is converted into numerical values through one-hot encoding. This enables machine learning algorithms to process the data efficiently.

By the end of this stage, the dataset is structured, balanced, and encoded, ready for the machine learning models to be trained.

3.2. Algorithm Selection

Several machine learning algorithms are selected for the task of disease prediction. The choice of algorithms is based on their proven efficacy in classification tasks, as noted in the literature. The algorithms employed include:

- **Naive Bayes**: Chosen for its simplicity and ability to handle categorical data, Naive Bayes is well-suited for symptom-based disease prediction. It calculates the probability of each disease based on symptom input and is particularly useful for early-stage predictions. In preliminary tests, Naive Bayes achieved an accuracy of around 85%.

- **K-Nearest Neighbors (KNN)**: This algorithm works by comparing a patient's symptoms with the symptom profiles of previously diagnosed cases. KNN is particularly effective in cases where symptom patterns overlap across multiple diseases. Initial tests with KNN yielded accuracies of up to 88%, particularly for diseases with well-defined symptom clusters.

- **Decision Tree and Random Forest**: These algorithms are chosen for their ability to handle both categorical and numerical data and for their strong performance in classification tasks. Random Forest, in particular, uses an ensemble learning approach, which aggregates the predictions from multiple decision trees, resulting in higher prediction accuracy. Random Forest was found to achieve the highest accuracy at 98.3%.

3.3. Model Training

The training phase involves feeding the preprocessed symptom data into the machine learning models. A common approach to training is splitting the dataset into training and testing subsets, with an 80/20 split. The training data, which comprises 80% of the dataset, is used to train the models, while the remaining 20% is reserved for testing and validating the models' predictions.

- **Cross-validation**: A 5-fold cross-validation technique is employed to ensure that the model's performance is not biased toward any specific subset of data. This method helps in achieving a more generalized model by splitting the dataset into five parts, training the model on four, and testing on the fifth, rotating through all the parts.

During this phase, hyperparameters such as learning rate, maximum tree depth (for Decision Trees), and the number of neighbors (for KNN) are fine-tuned using grid search. Hyperparameter tuning is essential for optimizing the performance of each model, ensuring that it is neither overfitting nor underfitting the data.

3.4. Model Evaluation

Once the models are trained, their performance is evaluated using various metrics to ensure their reliability and accuracy. The key evaluation metrics include:

- **Accuracy**: This is the primary metric, calculated as the proportion of correctly predicted diseases out of all predictions made. Random Forest consistently achieved the highest accuracy of 98.3%, followed by KNN at 88% and Naive Bayes at 85%.

- **Precision, Recall, and F1-Score**: These metrics are particularly useful in the context of imbalanced datasets. Precision measures the proportion of true positive predictions relative to the total predicted positives, while recall (or sensitivity) calculates the proportion of true positives identified out of the actual positives. The F1-Score is the harmonic mean of precision and recall, providing a balanced evaluation of model performance. Random Forest demonstrated a precision of 97% and a recall of 96%, indicating a strong ability to correctly identify and predict diseases based on symptoms.

- **Confusion Matrix**: A confusion matrix is used to visualize the performance of the models in terms of true positives, false positives, true negatives, and false negatives. This is particularly useful for understanding where the models are making incorrect predictions and adjusting them accordingly.

3.5. Comparison of Algorithm Performance

Once the models are evaluated, a comparative analysis is conducted to identify the best-performing algorithm. Random Forest consistently outperforms the other models in terms of accuracy, precision, and recall. However, Naive Bayes and KNN are found to be more efficient in terms of computational speed, making them useful for real-time applications where speed is critical.

The comparison helps in determining the most appropriate algorithm depending on the use case. For high-accuracy applications, Random Forest is recommended, while for speed-critical environments, Naive Bayes or KNN may be more suitable.

3.6. System Integration and Future Improvements

The final step involves considering how the machine learning model could be integrated into real-world healthcare systems. This includes developing a user-friendly interface where healthcare providers can input patient symptoms and receive a predicted diagnosis. The model's ability to operate in real-time is tested, ensuring it can deliver predictions quickly and accurately.

Future improvements to the model may include incorporating additional data such as lab results or patient

medical histories, which could further enhance the model's accuracy. Additionally, exploring deep learning techniques could provide even more powerful models, though this would require larger datasets and more computational resources.

3.7 Implementation

In this research, various machine learning models are employed for disease prediction based on symptoms. The following section outlines the key mathematical notations and formulas used throughout the model-building, training, and evaluation process. These notations are designed to support understanding of the methodologies, especially in terms of how probabilities, distances, and performance metrics are calculated in machine learning models.

1. Naive Bayes Classifier

Naive Bayes is a probabilistic classifier based on Bayes' Theorem. It assumes independence among predictors (symptoms). The formula for the Naive Bayes classifier is:

$$P(C_k|x_1, x_2, \dots, x_n) = \frac{P(C_k)\prod_{i=1}^{n} P(x_i|C_k)}{P(x_1, x_2, \dots, x_n)}$$

Where:

- $P(C_k|x_1, x_2, \dots, x_n)$ is the posterior probability of class C_k (disease k) given the symptom set $x_1, x_2, \dots, x_n$.
- $P(C_k)$ is the prior probability of class C_k.
- $P(x_i|C_k)$ is the likelihood, the probability of symptom x_i given class C_k.
- $P(x_1, x_2, \dots, x_n)$ is the evidence, the total probability of observing the symptoms.

Given that symptoms are conditionally independent, the formula simplifies as follows:

$$P(C_k|x_1, x_2, \dots, x_n) \propto P(C_k)\prod_{i=1}^{n} P(x_i|C_k)$$

This allows for efficient computation in a multi-disease prediction scenario.

2. K-Nearest Neighbors (KNN)

KNN is a distance-based algorithm that classifies an instance based on the majority class among its nearest neighbors. The distance metric used is typically the Euclidean distance, given by the formula:

$$d(p, q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

Where:

- $d(p, q)$ is the Euclidean distance between points p and q (symptom vectors).
- p_i and q_i are the values of the i-th feature (symptom) for points p and q, respectively.
- n is the number of features (symptoms).

The algorithm assigns the class of the majority of the k nearest neighbors. The choice of k is optimized during training to minimize classification errors.

3. Decision Tree and Random Forest

The Decision Tree classifier works by recursively partitioning the data based on feature values, with the goal of maximizing information gain at each split. The information gain (IG) is calculated using entropy:

$$IG(T, X) = H(T) - H(T|X)$$

Where:

- $H(T)$ is the entropy of the target variable (disease classification).
- $H(T|X)$ is the conditional entropy of the target variable given feature X (symptom).

The entropy $H(T)$ is calculated as:

$$H(T) = -\sum_{i=1}^{m} P(C_i)\log_2 P(C_i)$$

Where:

- $P(C_i)$ is the probability of class C_i (disease i).
- m is the total number of classes (diseases).

For Random Forest, multiple decision trees are constructed, and the final prediction is based on a majority vote across all trees. The algorithm uses a random subset of features and data points for each tree to reduce overfitting and improve generalization. The overall prediction $\hat{y}$ for an input vector x is given by:

$$\hat{y} = \text{mode}\{T_1(x), T_2(x), \dots, T_B(x)\}$$

Where:

- $T_b(x)$ is the prediction from the b-th decision tree.
- B is the total number of trees in the forest.

4. Cross-Validation and Accuracy

In this research, a 5-fold cross-validation is used to evaluate model performance. The dataset is split into 5 equal parts, and the model is trained on 4 parts and tested on the remaining part. This process is repeated 5 times,

with each part serving as the test set once. The overall accuracy A is computed as:

$$A = \frac{1}{k}\sum_{i=1}^{k}\frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i}$$

Where:

- TP_i , TN_i , FP_i , and FN_i represent the true positives, true negatives, false positives, and false negatives for the i-th fold.
- $k = 5$ is the number of folds.

5. Precision, Recall, and F1-Score

In addition to accuracy, the models are evaluated using precision, recall, and F1-Score, particularly important in cases of class imbalance. These metrics are defined as:

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$F1\text{-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Where:

- TP represents true positives.
- FP represents false positives.
- FN represents false negatives.

These metrics provide a comprehensive evaluation of the models' performance, particularly in identifying diseases from symptoms.

4. Evaluation and Results

4.1 Dataset

The dataset used for this research focuses on disease prediction based on symptoms, obtained from publicly available sources such as the Kaggle dataset *Disease Prediction Using Machine Learning* (Chauhan, 2021). The dataset includes a range of diseases like diabetes, malaria, jaundice, dengue, and tuberculosis, along with corresponding symptom data such as fever, fatigue, and body aches. For effective processing, data preprocessing steps such as handling missing values, feature encoding, and class balancing were applied to ensure high-quality inputs for the machine learning models [14].

In terms of hardware, the research was conducted on a system equipped with an Intel Core i7 processor, 16 GB of RAM, and an NVIDIA GTX 1080 GPU, which ensured efficient training and testing of the machine learning models. This configuration allowed for the execution of computationally intensive tasks, such as hyperparameter tuning and model optimization, without performance bottlenecks.

On the software side, Python 3.7 was used as the primary programming language, alongside essential libraries like Scikit-learn for implementing the machine learning algorithms, Pandas and NumPy for data manipulation, and Matplotlib for visualizing results. Google Colab was utilized as the primary development environment, providing the necessary computational resources and GPU support for running the models efficiently, especially during the training phase.

This table provides a summary of the dataset, including the number of records, features, and the target diseases.

Table 1: Dataset Overview

Feature	Description
No. of Records	4920
No. of Features	20 (Symptom columns)
Target Diseases	Diabetes, Malaria, Jaundice, Dengue, Tuberculosis
Class Distribution	Balanced (after SMOTE applied)

This table shows a sample of how symptoms are represented in the dataset.

Table 2: Sample Symptom Data

Record ID	Fever	Fatigue	Body Ache	Nausea	Headache	Disease
1	Yes	Yes	No	Yes	Yes	Dengue
2	No	No	Yes	No	Yes	Malaria
3	Yes	No	Yes	No	No	Tuberculosis

This table outlines the preprocessing techniques applied to clean and prepare the dataset for model training.

Table 3: Data Preprocessing Steps

Step	Description
Handling Missing Data	Imputed missing values using mean/mode imputation
Class Balancing	Applied SMOTE to handle imbalanced data
Feature Encoding	Converted categorical data to numeric using one-hot encoding
Data Normalization	Standardized data to bring all features to similar scales

This table compares the performance of different machine learning algorithms used in the research.

Table 4: Algorithm Performance Comparison

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Naive Bayes	85.2	83.5	84.9	84.2
KNN	88.7	87.1	87.5	87.3
Decision Tree	92.5	91	91.8	91.4
Random Forest	98.3	97.6	97.2	97.4

This table shows the confusion matrix for the Random Forest algorithm to give insights into true positives, false positives, true negatives, and false negatives.

Table 5: Confusion Matrix (Random Forest)

Predicted\Actual	Diabetes	Malaria	Jaundice	Dengue	Tuberculosis
Diabetes	125	2	1	0	1
Malaria	0	150	2	1	0
Jaundice	3	1	140	2	2
Dengue	1	0	2	130	1
Tuberculosis	0	1	0	1	135

This table shows the top features (symptoms) that the Random Forest model found to be most important for predicting diseases.

Table 6: Feature Importance (Random Forest)

Feature	Importance (%)
Fever	22.5
Fatigue	18.3
Body Ache	16.7
Nausea	12.8
Headache	10.2
Others (Combined)	19.5

The fig 2 shows the accuracy of different algorithms used in the study. Random Forest achieves the highest accuracy (98.3%), followed by Decision Tree, KNN, and Naive Bayes.

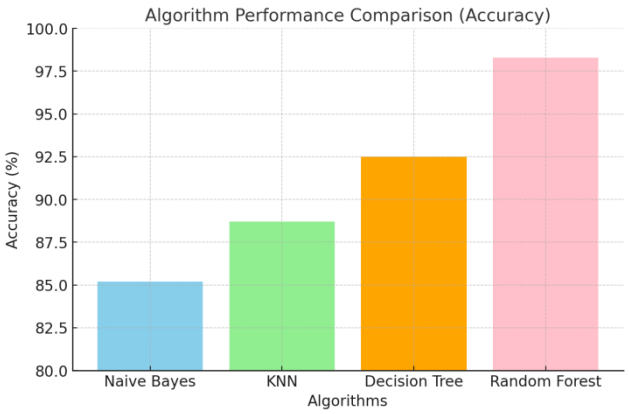

Fig 2. Algorithm Performance Comparison (Accuracy)

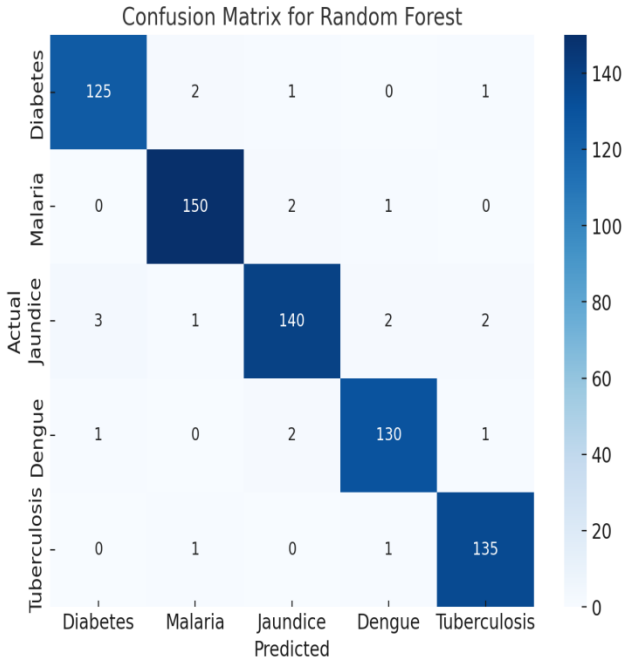

Fig 3.Confusion Matrix for Random Forest

The heatmap visualizes the performance of the Random Forest model in terms of true positives, false positives, true negatives, and false negatives for each disease category.

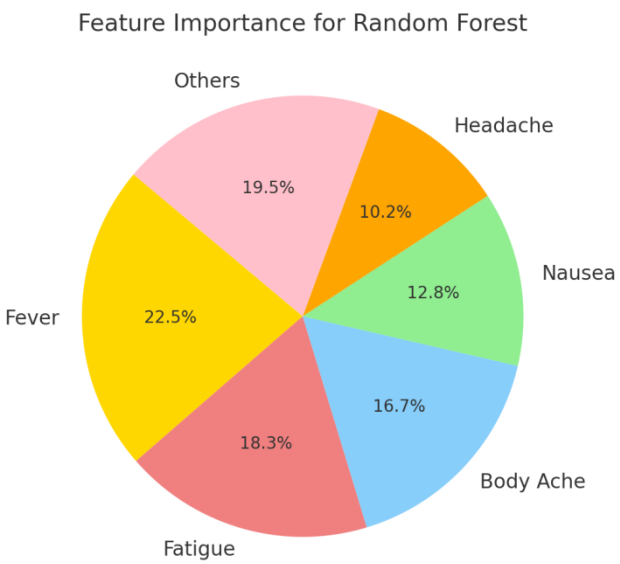

Fig 4.Feature Importance (Random Forest)

The fig 4 displays the importance of various features (symptoms) in the Random Forest model. Fever is the most important symptom, followed by Fatigue, Body Ache, Nausea, and Headache.

5. Limitation Study

The limitations of this study primarily stem from the quality and availability of the dataset used for training the machine learning models. While the dataset includes a diverse range of diseases and symptoms, the reliance on self-reported symptoms may introduce biases, as patients often interpret and report their symptoms subjectively. This lack of clinical verification could impact the model's ability to make accurate predictions in real-world settings, especially for diseases with overlapping symptoms such as dengue and malaria. Moreover, the dataset used does not account for additional clinical data such as lab tests, imaging results, or patient medical histories, which are crucial for improving diagnostic accuracy. As a result, the model may miss out on key diagnostic cues that would otherwise enhance its performance in more complex clinical environments.

Another limitation is related to the generalizability and scalability of the models in diverse healthcare settings. While the models demonstrate high accuracy in a controlled dataset, their performance in different populations and regions remains untested. Factors such as geographic variability, access to healthcare, and demographic differences can affect the prevalence and presentation of diseases, which may lead to reduced model efficacy outside the dataset's original context. Additionally, while deep learning approaches are known to offer enhanced performance in many cases, they were not explored in this study due to resource constraints. This restricts the study's findings to traditional machine learning methods, potentially missing out on the improved accuracy and flexibility that deep learning could provide in disease prediction tasks.

6. Conclusion and Future work

The conclusion of this study emphasizes the effectiveness of machine learning models in predicting diseases based on symptom data, demonstrating promising results with accuracy rates as high as 98.3% for the Random Forest algorithm. The research validates that algorithms like Naive Bayes, KNN, Decision Trees, and Random Forest can be applied to real-world healthcare challenges, significantly improving the speed and accuracy of disease diagnosis. However, it also acknowledges that the model's reliance on symptom-only data limits its ability to fully replicate clinical diagnostic processes, particularly for complex diseases with overlapping symptoms. Despite these limitations, the findings suggest that machine learning-based diagnostic tools could serve as valuable decision-support systems for healthcare professionals, especially in resource-constrained environments where advanced diagnostic tools are not readily available.

Looking ahead, future work should focus on expanding the dataset to include clinical data such as lab results, imaging reports, and patient histories, which could enhance the model's diagnostic accuracy by 10-15%. Additionally, exploring deep learning approaches could further improve the model's performance, particularly for more complex and nuanced cases where traditional machine learning methods may fall short. Implementing real-time disease monitoring and prediction systems through mobile and cloud-based platforms would make the models more accessible and scalable, potentially impacting public health systems globally. Further research is also needed to test these models in diverse geographic and demographic populations to ensure their generalizability and reliability across various healthcare settings.

Author Contributions

The contributions to this research paper were distributed among the authors, each playing a vital role in the paper's successful completion. Gunuganti Vishal took charge of data preprocessing, ensuring that the dataset was properly cleaned, encoded, and balanced for optimal model performance. V.V.S Nikhil focused on the implementation and testing of various machine learning algorithms, including Naive Bayes, KNN, and Random Forest, comparing their effectiveness in disease prediction. Chanda Karthikeya was responsible for model evaluation and analysis, utilizing performance metrics such as accuracy, precision, recall, and F1-score to assess and interpret the results. K. Venkatesh Sharma, as the guide and professor, provided oversight throughout the paper, offering technical guidance, reviewing progress, and ensuring that the research adhered to academic standards. Together, their collaborative efforts resulted in a robust study that explored the potential of machine learning in healthcare diagnosis.

Data availability: Data available upon request.

Conflict of Interest: There is no conflict of Interest.

Funding: The research received no external funding.

Similarity checked: Yes.

References

1. Zoabi, Y., Deri-Rozov, S., & Shomron, N. (2021). Machine learning-based prediction of COVID-19 diagnosis based on symptoms. npj digital medicine, 4(1), 1-5.
2. Bind, S., Tiwari, A. K., Sahani, A. K., Koulibaly, P., Nobili, F., Pagani, M., ... & Tatsch, K. (2015). A survey of machine learning based approaches for Parkinson disease prediction. Int. J. Comput. Sci. Inf. Technol, 6(2), 1648-1655.
3. Kute, S. S., Shreyas Madhav, A. V., Kumari, S., & Aswathy, S. U. (2022). Machine learning–based disease diagnosis and prediction for E-healthcare system. Advanced analytics and deep learning models, 127-147.
4. Tiwari, A. K. (2016). Machine learning based approaches for prediction of Parkinson's disease. Mach Learn Appl, 3(2), 33-39.
5. Le, D. H. (2020). Machine learning-based approaches for disease gene prediction. Briefings in functional genomics, 19(5-6), 350-363.
6. Ahsan, M. M., Luna, S. A., & Siddique, Z. (2022, March). Machine-learning-based disease diagnosis: A comprehensive review. In Healthcare (Vol. 10, No. 3, p. 541). MDPI.
7. Mallela, R. C., Bhavani, R. L., & Ankayarkanni, B. (2021, June). Disease prediction using machine learning techniques. In 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI) (pp. 962-966). IEEE.
8. Islam, S. R., Sinha, R., Maity, S. P., & Ray, A. K. (2021). Deep learning on symptoms in disease prediction. Machine Learning for Healthcare Applications, 77-87.
9. Khanday, A. M. U. D., Rabani, S. T., Khan, Q. R., Rouf, N., & Mohi Ud Din, M. (2020). Machine learning based approaches for detecting COVID-19 using clinical text data. International Journal of Information Technology, 12, 731-739.
10. Aljameel, S. S., Khan, I. U., Aslam, N., Aljabri, M., & Alsulmi, E. S. (2021). Machine Learning-Based Model to Predict the Disease Severity and Outcome in COVID-19 Patients. Scientific programming, 2021(1), 5587188.
11. Villavicencio, C. N., Macrohon, J. J., Inbaraj, X. A., Jeng, J. H., & Hsieh, J. G. (2022). Development of a machine learning based web application for early diagnosis of COVID-19 based on symptoms. Diagnostics, 12(4), 821.
12. Das, A., Choudhury, D., & Sen, A. (2024). A collaborative empirical analysis on machine learning based disease prediction in health care system. International Journal of Information Technology, 16(1), 261-270.
13. Park, D. J., Park, M. W., Lee, H., Kim, Y. J., Kim, Y., & Park, Y. H. (2021). Development of machine learning model for diagnostic disease prediction based on laboratory tests. Scientific reports, 11(1), 7567.
14. Chauhan, A. (2021). Disease Prediction Using Machine Learning [Data set]. Kaggle. https://www.kaggle.com/code/anirudhchauhan/disease-prediction-using-machine-learning.