IJNRD.ORG ISSN : 2456-4184

INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG

An International Open Access, Peer-reviewed, Refereed Journal

A NOVEL MACHINE LEARNING APPROACH TO PREDICT MULTIPLE DISEASES

Aniket Kumar, Abhay Sharma, Aayush Sharma, Aditya Bansal
Student, Student, Student, Student
Department of (CSE)
Meerut Institute of Engineering and Technology, Meerut, India

ABSTRACT:

In recent years, machine learning (ML) has shown great promise in revolutionizing healthcare by improving the accuracy and efficiency of disease detection and diagnosis. This research explores the development of a multi-disease detection system that leverages various ML techniques, such as Decision Trees and Random Forests. The main goal is to uncover patterns and relationships between a set of symptoms and four common diseases: diabetes, heart disease, Parkinson's disease, and other symptom-based predictions.

By applying these algorithms, we aim to build a powerful model that predicts the likelihood of disease & also provides valuable insights into symptoms may be interconnected. The dataset used in this study includes a broad range of patient symptoms and outcomes, allowing for a thorough analysis. We rigorously test the model's performance using key evaluation to ensure its reliability. Initial results show that machine learning models can significantly enhance diagnostic accuracy, offering valuable support to healthcare professionals in making clinical decisions. This research contributes to the expanding field of medical informatics and underscores the transformative potential of ML in healthcare. Looking ahead, future work will focus on incorporating additional algorithms and expanding the dataset to further refine predictive capabilities and enhance the system's clinical relevance.

KEYWORDS: Machine Learning, Disease Prediction, HealthCare, Logistic Regression, Random Forest, Decision Tree,

Index

1 INTRODUCTION:

General Overview of the Topic

The rapid growth of technology and data analytics has transformed many industries, and healthcare is one of the sectors that has experienced the most significant impact. The integration of machine learning (ML) into healthcare offers new and exciting possibilities for improving disease detection, diagnosis, and management. As chronic diseases continue to rise worldwide, there is a pressing need for more efficient and accurate diagnostic tools to help healthcare providers make better, more informed decisions. These algorithms can reveal complex patterns and relationships within large datasets—patterns that may be too subtle for traditional methods to detect. By harnessing ML, healthcare professionals can improve their ability to diagnose diseases based on a patient's symptoms, medical history, and other relevant factors.

In this study, we focus on using machine learning to detect several prevalent diseases, specifically diabetes, heart disease, and Parkinson's disease. Each of these conditions presents unique challenges, both in terms of diagnosis and management, and often requires a deep understanding of symptoms and patient history. We use algorithms like Decision Trees and Random Forests to build a predictive model that can assess the likelihood of a disease based on the symptoms a patient presents.

A key part of this research is ensuring that the model is reliable. Through rigorous testing and validation.

Ultimately, this research aims to bridge the gap between cutting-edge machine learning technology and real-world healthcare applications. By creating a robust system for multi-disease detection, we hope to improve diagnostic accuracy, help healthcare providers deliver better care, and ultimately improve patient outcomes.

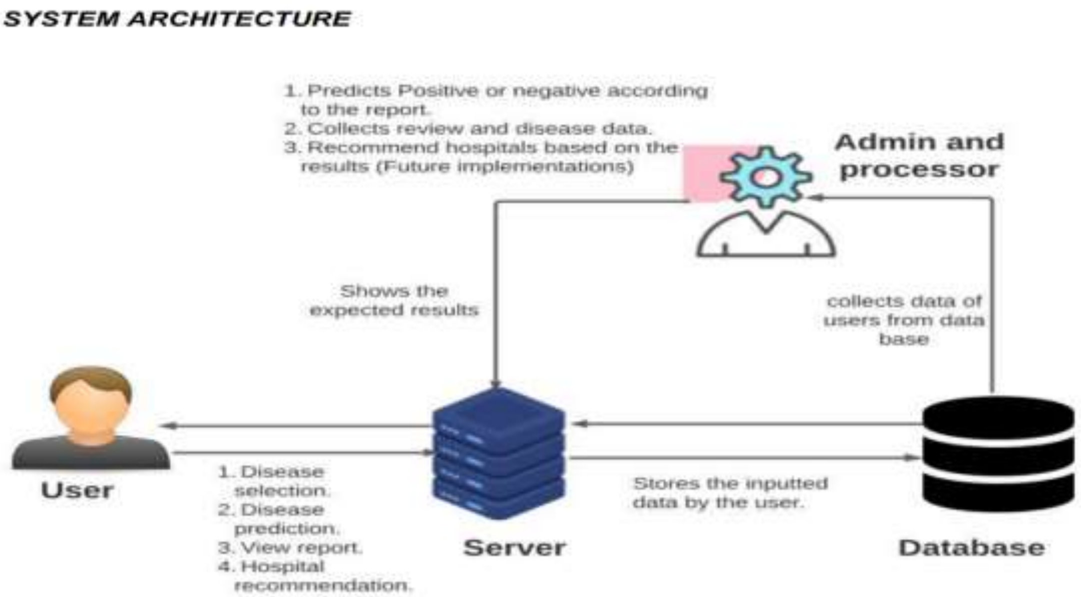

SYSTEM ARCHITECTURE
1. Predicts Positive or negative according to the report.
2. Collects review and disease data.
3. Recommend hospitals based on the results (Future implementations)
Admin and processor
Shows the expected results
collects data of users from data base
User
Server
Database
1. Disease selection.
2. Disease prediction.
3. View report.
4. Hospital recommendation.
Stores the inputted data by the user.

Fig 4.5: System Architecture.

2. Major Diseases

2.1 Diabetes Disease

Overview: Diabetes is a chronic condition that affects how the body processes blood sugar (glucose). It primarily comes in two forms: Type 1 diabetes, where the body doesn't produce insulin, and Type 2 diabetes, where the body doesn't use insulin properly.

Symptoms:

- Extreme fatigue
- Blurred vision
- Slow-healing sores or frequent infections

Impact: Without proper management, diabetes can lead to serious complications such as heart disease, nerve damage, kidney problems, and vision loss. Early diagnosis and intervention are key to preventing long-term health issues.

2.2 Heart Disease

Overview: Heart disease encompasses a variety of conditions that impact the heart’s structure and how it functions. This includes issues like **coronary artery disease**, where the blood vessels supplying the heart become narrowed.

Symptoms:

- Chest pain
- Breath problem,
- Unbalanced heartbeat

Impact: Heart disease is one of the leading causes of death around the world. Catching it early and managing it effectively are key to lowering the chances of more serious issues like heart attacks, strokes.

2.3 Parkinson’s Disease

Overview: It occurs when the brain loses its ability to produce dopamine, a chemical that helps control muscle movements.

Symptoms:

- Tremors or shaking, especially in the hands
- Muscle pain,
- Changes in speech and handwriting

Impact: Parkinson's disease can severely affect a person's quality of life, leading to mobility problems, cognitive decline, and a range of non-motor symptoms. While it is not curable, early diagnosis and ongoing treatment can help manage symptoms and improve the patient's ability to live independently.

2.4 Symptom Based Prediction

Overview: Symptom-based prediction uses machine learning algorithms to analyze reported symptoms and predict possible diseases. Users input symptoms, and the system provides likely diagnoses based on medical data and pattern recognition.

Impact: This approach enables early detection of diseases, helping users seek timely medical advice and reducing the risk of complications. It also supports healthcare professionals by offering quick, data-driven preliminary assessments.

3. LITERATURE REVIEW:

3.1 Diabetes Prediction

The use of ML in predicting diabetes, particularly Type 2 diabetes, has garnered significant attention. **Sharma et al. (2019)** developed a predictive model using Decision Trees and Random Forests to identify individuals at risk of developing Type 2 diabetes. Their model achieved over 85% accuracy, showcasing how effectively ML can analyze patient data to forecast disease onset. Similarly, Kumar and Kumar (2020) explored the use of support vector machines (SVMs) for diabetes prediction, achieving similar results. These findings support the growing belief that ML models are powerful tools for diabetes risk assessment, helping doctors intervene earlier and more effectively.

3.2 Heart Disease Prediction

In a study by **Dua and Graff (2019),** various ML algorithms, including Logistic Regression and Random Forests, were used to analyse the UCI Heart Disease dataset. The results indicated that Random Forests performed particularly well, achieving an accuracy of around 90%. This demonstrates the ability of ML models to identify the risk factors associated with heart disease. Additionally, Alzubaidi et al. (2020) applied ensemble learning techniques to further enhance prediction accuracy.

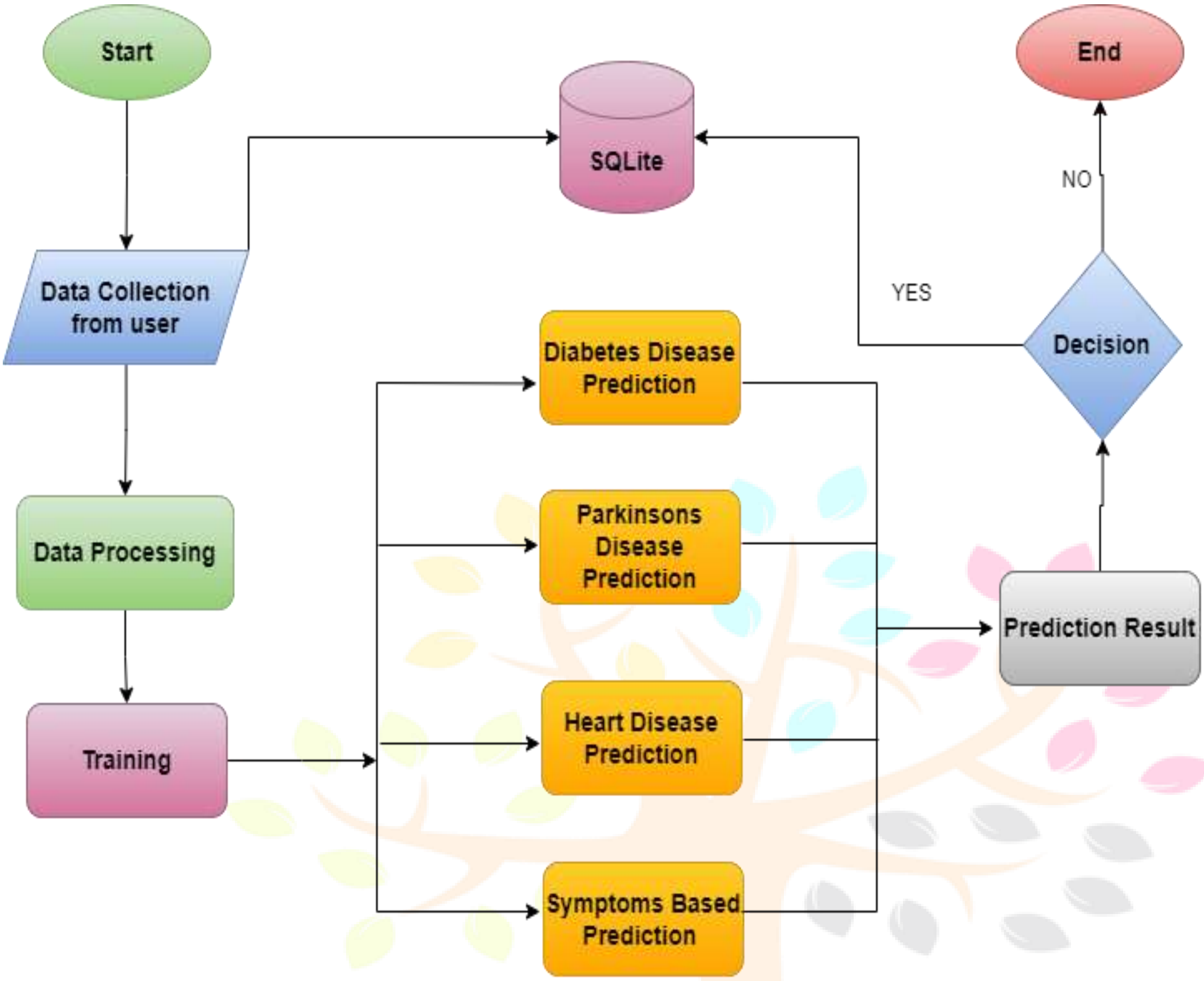
3.3 Parkinson's Disease Detection

Parkinson's disease poses unique challenges due to its progressive nature and complex symptomatology. However, recent advancements in ML are helping to improve early detection. **Gonzalez et al. (2021)** explored the use of deep learning techniques to analyze voice and gait patterns in patients with Parkinson's disease. Their model demonstrated high sensitivity and specificity, indicating that ML can detect subtle changes that might otherwise be overlooked by traditional diagnostic methods. Moreover, Santos et al. (2022) combined feature selection techniques with ML algorithms to improve the accuracy of predicting Parkinson's disease using clinical data.

4. WORKING:

The process begins with **data collection**, where a comprehensive dataset is gathered, including medical records, symptoms, and diagnostic results for a variety of diseases. This dataset forms the foundation for all subsequent analysis.
The next step is **data preprocessing**, which involves cleaning the data and standardizing it to ensure consistency. This is an essential process to remove any noise or errors in the data, making it ready for more precise analysis. During this stage, we also handle missing values, correct inconsistencies, and normalize the data, ensuring it is in the best possible shape for training machine learning models.

Following preprocessing, **feature extraction and selection** are performed. This step focuses on identifying the most important attributes or features that are most relevant for predicting the presence of a disease. By carefully selecting these key


Start
SQLite
End
NO
Data Collection from user
YES
Decision
Diabetes Disease Prediction
Parkinsons Disease Prediction
Data Processing
Prediction Result
Training
Heart Disease Prediction
Symptoms Based Prediction

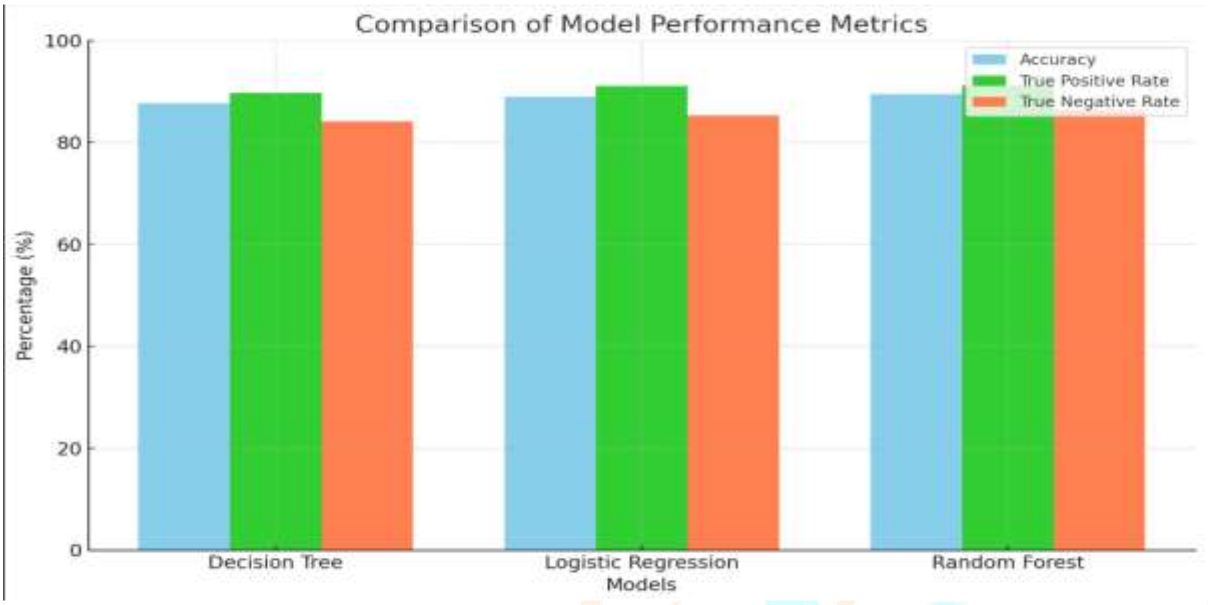
features, we improve the accuracy of the model, ensuring that it can make precise predictions based on the most impactful data points.
Once the data is clean and the most relevant features have been selected, we move on to **model development**. These algorithms are designed to learn from the data and make predictions about disease presence based on the input symptoms.

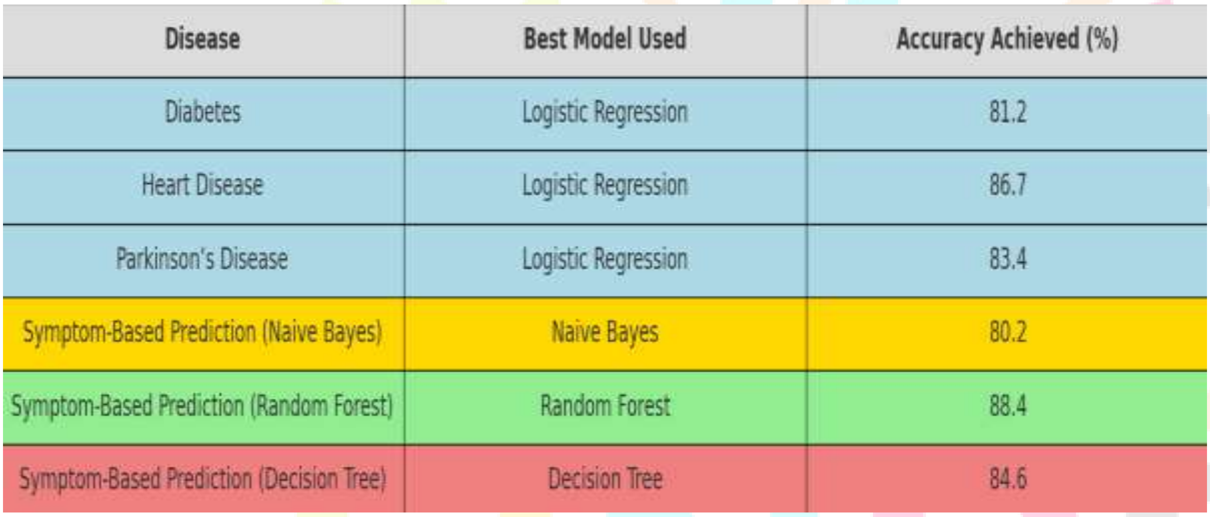
5. RESULT AND DISCUSSION:

We tested 5 different ml models to predict diseases based on an available dataset. Out of these, six models achieved an accuracy of 50% or higher. The Logistic Regression model performed the best, achieving an impressive accuracy of (81-86) %. By allowing both smaller and larger values during training, the Weighted KNN was able to adapt and make more accurate predictions.
The dataset was analysed based on key factors such as gender, age group, and symptoms, providing a well-rounded input for each model. However, not all models performed equally well. The Random Forestt Tree model struggled significantly, with an accuracy of 88.4%.

When it came to the Naïve Bayes models, both Gaussian Naïve Bayes and Decision Tree recorded an accuracy of above 80%, which is on the lower end of the spectrum. On a more positive note, the Naïve Bayes model showed a much better result, with an accuracy of 80.2%. while the Decision Tree model performed poorly, with just 4.3% accuracy. Interestingly, the Decision Tree model showed a solid performance, reaching an accuracy of 84.6%.

These findings highlight that the choice of model is crucial for disease prediction, and while some models like the Weighted KNN were highly effective, others struggled to deliver reliable results. This variation in performance helps guide future research into selecting the best machine learning.





Disease	Best Model Used	Accuracy Achieved (%)
Diabetes	Logistic Regression	81.2
Heart Disease	Logistic Regression	86.7
Parkinson's Disease	Logistic Regression	83.4
Symptom-Based Prediction (Naive Bayes)	Naive Bayes	80.2
Symptom-Based Prediction (Random Forest)	Random Forest	88.4
Symptom-Based Prediction (Decision Tree)	Decision Tree	84.6

Model Accuracy for Disease Prediction (Colored by Model Type)

6. CHALLENGES AND FUTURE DISCUSSION

While the results from these studies are promising, there are still several challenges in applying ML for disease detection. Key issues include the quality of data, the interpretability of ML models, and the need for large, diverse datasets to ensure the models are accurate and generalizable. Many current models are trained on limited datasets, which can affect their reliability when applied to different populations. To address these challenges, future research

should focus on integrating real-world clinical data, improving model transparency, and exploring advanced techniques like deep learning and ensemble methods. These approaches may hold the key to overcoming some of the current limitations and making ML-driven diagnostic tools more reliable and accessible in clinical settings.

7. REFERENCE:

[1] UCI Machine Learning Repository. (n.d.). Retrieved from UCI ML Repository.
[2] Khan, M. A., & Alzahrani, A. I. (2020). A comprehensive review of machine learning techniques for healthcare applications. *Journal of Healthcare Engineering, 2020*, 1-20.
[3] Pang, Z., & Wang, Q. (2021). Machine learning in disease prediction: A systematic review. *Artificial Intelligence in Medicine, 113*, 101986.
[4] Zhang, Y., & Wang, H. (2019). An overview of machine learning methods for healthcare applications. *Journal of Healthcare Informatics Research, 3*(4), 1-21.
[5] Chawla, N. V., & Bowyer, K. W. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research.*
[6] Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5-32.
[7] Multi Disease Prediction Using Data Mining Techniques" K. Gomathi, Dr. D. Shanmuga Priya (2017).