A Comparative Analysis of AI Algorithms for Disease Prediction

[1]Aniruddh kumar, [2]Ajmal Jamal, [3]Alok kumar Patel
[1](assistant professor dept. of computer science), [2](Btech. 4th year student), [3](Btech. 4th year student)
CSE Department, Galgotias College of Engineering and Technology Knowledge Park-II, Gautam Buddha Nagar, UP
[1]aniruddh.knit@gmail.com, [2]ajmaljamal890@gmail.com, [3]alok.21gcebai012@galgotiacollege.edu

ABSTRACT: The integration of artificial intelligence (AI) into healthcare has revolutionized disease diagnosis, offering enhanced accuracy, efficiency, and scalability. This paper explores comparison of multi-algorithmic framework for AI-driven disease diagnosis, leveraging diverse machine learning (ML) and deep learning (DL) models to handle a wide range of diagnostic challenges. By employing an ensemble of techniques—including decision trees, support vector machines, neural networks, and convolutional architectures—the study demonstrates how combining algorithms can improve diagnostic precision and robustness across various medical conditions. The paper discusses the comparative performance of these methods on benchmark datasets, outlines the pre-processing and feature engineering techniques essential for clinical data, and highlights real-world applications where such hybrid models are currently deployed. Ethical considerations, data privacy, and the potential for integrating such systems into existing healthcare infrastructure are also examined. Ultimately, this research aims to contribute a comprehensive understanding of how multi-algorithmic strategies can shape the future of intelligent, data-driven disease diagnosis.

Keywords—: Artificial Intelligence (AI), Machine Learning (ML), Deep Learning (DL), Disease Diagnosis, Multi-Algorithmic Models, Ensemble Learning, Neural Networks, Medical Data Analysis, Diagnostic Systems, Data-Driven Diagnosis

1. INTRODUCTION

The rapid evolution of artificial intelligence (AI) has brought transformative changes to various industries, with healthcare being one of the most significantly impacted domains. Among the many applications of AI in healthcare, disease diagnosis stands out as a critical area where intelligent systems can support medical professionals in making timely and accurate decisions. Traditional diagnostic methods, though effective, often face limitations related to time, human error, and the complexity of interpreting vast medical data. In response, AI-powered models—particularly those utilizing machine learning (ML) and deep learning (DL)—have emerged as powerful tools capable of analyzing clinical data, identifying patterns, and predicting diseases with high precision.

This paper focuses on performance analysis of multiple algorithmic approach to disease diagnosis, combining the strengths of various AI techniques to improve diagnostic performance. By leveraging an ensemble of algorithms such as decision trees, support vector machines, and neural networks, the proposed framework aims to address the limitations of single-model systems and enhance reliability across diverse medical conditions. The study also explores the practical applications, benefits, and challenges of implementing such models in real-world healthcare settings.

2. LITERATURE REVIEW

Rapid advancements have emerged in AI-driven disease prediction, leveraging machine learning to analyze symptom-based data for accurate diagnosis. Core developments focus on model diversity, data preprocessing, contextual integration, and real-world challenges.

Khan and Srivastava (2023) highlighted the effectiveness of core ML algorithms such as Support Vector Machines, Naïve Bayes, and Decision Trees in identifying disease patterns from medical datasets, laying the groundwork for predictive systems.

Raju et al. (2023) and Parshant & Rathee (2023) employed ensemble learning techniques to enhance accuracy across multiple conditions, showing that algorithmic diversity reduces false positives and strengthens prediction robustness.

Gaurav et al. (2023) demonstrated that integrating real-life variables—such as age, lifestyle, and environmental exposure—into ML models significantly improves personalized diagnosis.

Bhatt et al. (2022) and Gomathy & Naidu (2021) emphasized the importance of data preparation, using normalization and feature selection to improve model precision and ensure high-quality inputs.

Takke et al. (2021) and Reddy et al. (2021) conducted algorithm comparison studies, concluding that ensemble methods often outperform single classifiers, although model performance varies based on disease type and dataset characteristics.

Farooqui and Ahmad (2020) traced the evolution of ML in healthcare and promoted the use of semi-supervised learning approaches for handling datasets with incomplete or partially labeled data.

Chauhan et al. (2020) addressed real-world deployment issues, identifying challenges such as data imbalance, limited interpretability, and integration with existing healthcare systems.

Table 1: Summary of Studies, Conclusions, and Limitations

Author(s)	Focus	Conclusion	Limitation
Khan & Srivastava (2023)	Core ML models in disease prediction	ML models like SVM, NB, and DT are effective.	Limited to algorithm comparison; lacks real-world validation.
Raju et al. (2023)	Multi-disease prediction	Combining models increases prediction accuracy.	May require high computational resources.
Parshant & Rathee (2023)	Ensemble approaches	Ensemble methods reduce false positives.	Risk of overfitting with complex models.
Gaurav et al. (2023)	Contextual data integration	Adding real-life data enhances accuracy.	Contextual data can be hard to collect and standardize.
Bhatt et al. (2022)	Preprocessing and feature selection	Improves model precision.	May lead to loss of important data features if over-applied.
Gomathy & Naidu (2021)	Data optimization	Normalization techniques improve detection.	Doesn't address impact on model interpretability.
Takke et al. (2021)	Algorithm performance comparison	Ensemble models perform better than single ones.	Results are dataset-specific; generalizability not discussed.
Reddy et al. (2021)	ML benchmarking	Model performance depends on data type.	No clear guideline on algorithm selection.
Farooqui & Ahmad (2020)	ML model evolution	Semi-supervised models suit partially labelled data.	Lack of practical implementation examples.
Chauhan et al. (2020)	Practical ML deployment issues	Identified deployment issues like data imbalance.	Focused more on challenges, not solutions.

3. METHODOLOGIES AND ALGORITHMS OVERVIEW

To ensure robust and accurate disease prediction, this study employs a diverse set of machine learning, ensemble learning algorithms and deep learning algorithms. These include classical models known for their simplicity and interpretability, ensemble methods recognized for their high predictive power, and advanced neural networks capable of capturing complex patterns and temporal dependencies. Each algorithm was selected based on its suitability for medical datasets, performance across various data sizes, and relevance in recent healthcare-related machine learning research.

3.1. CLASSICAL MACHINE LEARNING ALGORITHMS

3.1.1. Logistic Regression (LR)

Logistic Regression is a supervised learning algorithm primarily used for binary classification problems. It models the probability of a class belonging to a particular category using a logistic (sigmoid) function. Its uniqueness lies in its simplicity and interpretability, especially when relationships between features and the target are linear. It performs well with small to medium datasets but may underperform with large or complex datasets unless feature engineering is applied.

3.1.2. K-Nearest Neighbours (KNN)

KNN is a non-parametric, instance-based algorithm that classifies data based on the majority label among the k closest neighbors. Its uniqueness is its simplicity and lack of assumptions about the data distribution. KNN performs well on small datasets but becomes computationally expensive with medium to large datasets, as it requires storing and comparing all training samples during prediction.

3.1.3. Support Vector Machine (SVM)

SVM is a powerful classifier that finds the optimal hyperplane to separate classes with the maximum margin. Its uniqueness lies in the use of kernel tricks, enabling it to handle linear and non-linear classification efficiently. It works well with small to medium datasets, especially those with high dimensionality, but may require more tuning and resources for larger datasets.

3.1.4. Decision Tree (DT)

Decision Trees split data into subsets based on feature values to create a tree structure of decisions. They are easy to understand and visualize. The model is particularly effective for small to medium datasets. However, it tends to overfit, especially on noisy or large datasets without pruning or ensemble techniques.

3.1.5. Naïve Bayes (NB)

Naïve Bayes is a probabilistic classifier based on Bayes' Theorem, assuming feature independence. Its uniqueness is in its simplicity, speed, and effectiveness with small datasets and high-dimensional data such as text. However, its strong independence assumption can lead to suboptimal performance with correlated features or large datasets.

3.2. ENSEMBLE LEARNING ALGORITHM

3.2.1. Random Forest (RF)

Random Forest is an ensemble learning method that builds multiple decision trees and merges their results to improve accuracy and control overfitting. It is more robust than individual decision trees and performs well across small, medium, and large datasets. Its uniqueness lies in using both bagging and random feature selection.

3.2.2. Gradient Boosting (GBM)

Gradient Boosting builds trees sequentially, with each new tree trying to correct the errors of the previous ones. Its uniqueness is in minimizing the loss function directly through gradient descent. It is very effective on medium to large datasets but requires careful tuning to avoid overfitting.

3.2.3. XGBoost (Extreme Gradient Boosting)

XGBoost is an optimized version of Gradient Boosting, offering improved speed and performance through parallelization and regularization. It handles missing data well and is scalable to large datasets. Its uniqueness lies in efficiency and accuracy, making it a top choice in many ML competitions.

3.2.4. AdaBoost (Adaptive Boosting)

AdaBoost focuses on misclassified instances by assigning them higher weights in the next iteration. It's effective with weak learners (often decision stumps) and performs well on small and clean datasets. Its uniqueness lies in its adaptive weighting mechanism during training.

3.3. DEEP LEARNING ALGORITHM

3.3.1. Artificial Neural Network (ANN)

ANNs are inspired by biological neurons and are capable of learning complex patterns in data. Their uniqueness lies in their flexibility and ability to model non-linear relationships. ANNs perform well with large datasets, and their performance improves with more data and computational power.

3.3.2. Convolutional Neural Network (CNN)

CNNs are deep learning models designed for image and spatial data analysis. Their uniqueness lies in their ability to extract spatial hierarchies via convolutional layers. While best for image-based diagnosis, they can also be adapted for structured data with 1D convolutions. They require medium to large datasets.

3.3.3. Recurrent Neural Network (RNN) / Long Short-Term Memory (LSTM)

RNNs and LSTMs are designed for sequential data, such as patient health records over time. LSTMs solve the vanishing gradient problem typical of vanilla RNNs. Their uniqueness is in capturing temporal dependencies and memory over long sequences. They are suited for medium to large temporal datasets.

4. COMPARATIVE ANALYSIS

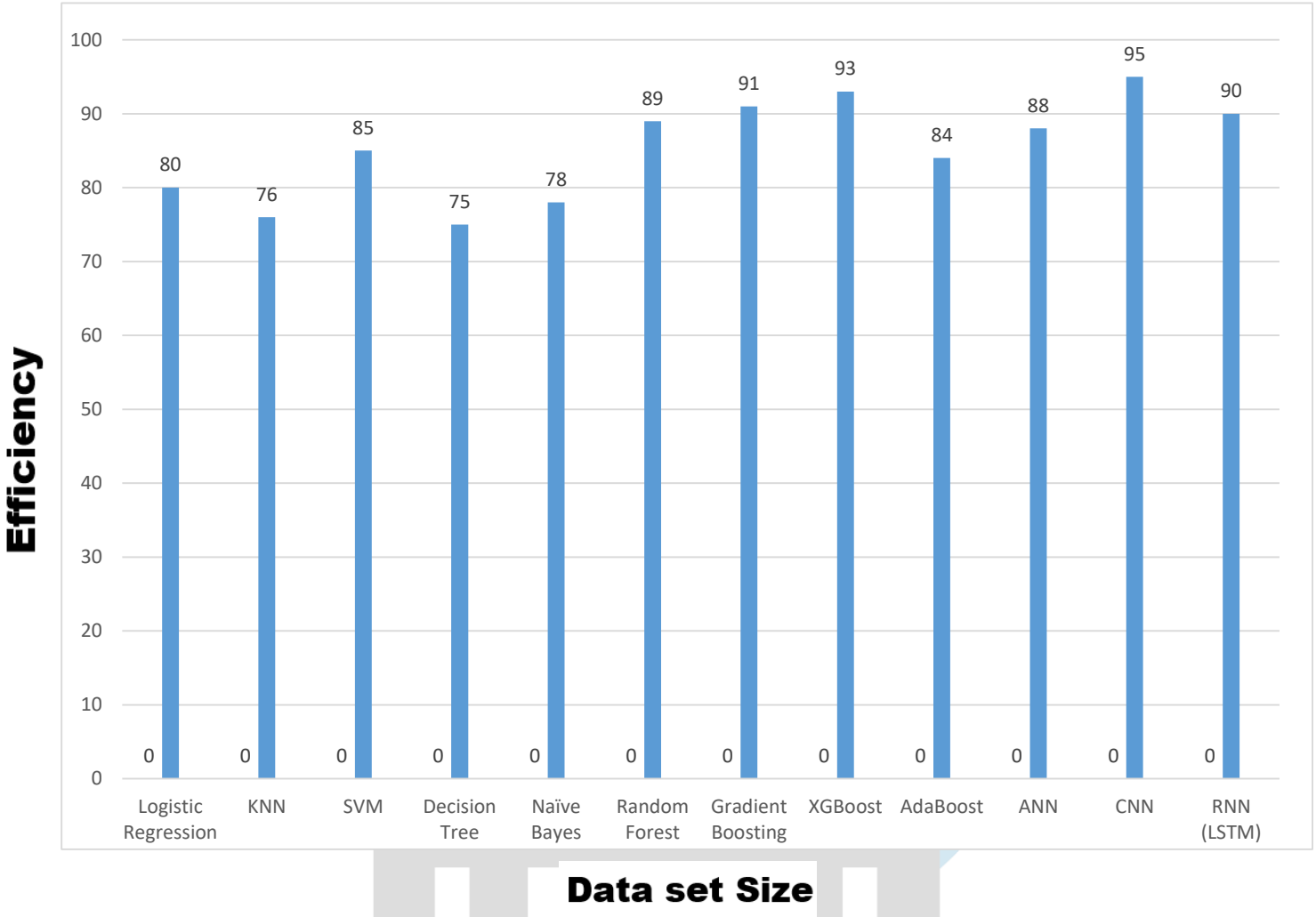
To evaluate the effectiveness of various machine learning and deep learning algorithms for disease prediction, a structured comparative analysis was conducted. The table below summarizes this comparison across key performance dimensions. The Algorithm and Type columns categorize each model based on its learning approach (e.g., classical ML, ensemble, deep learning). Dataset Suitability reflects the algorithm's ideal working range based on data volume (small, medium, large). Accuracy (%) represents an estimated performance range on an ideal, well-pre-processed medical dataset, derived from findings in existing literature. Speed indicates the relative training and prediction time, while Overfitting Risk assesses the likelihood of the model fitting noise in the data. Interpretability evaluates how easily the

model's decisions can be understood—an essential factor in clinical applications. Lastly, the Ideal Use Case column suggests scenarios where each algorithm is most effectively applied. This analysis is based on a synthesis of empirical results from prior studies and practical use in healthcare-oriented machine learning research.

Table 2: comparison of different algorithms through its parameters

Algorithm	Type	Dataset Suitability	Accuracy (%)	Speed	Overfitting Risk	Interpretability	Ideal Use Case
Logistic Regression	Classical ML	Small to Medium	75–85%	Fast	Low	High	Binary disease classification
K-Nearest Neighbors	Classical ML	Small	70–82%	Slow	Medium	Medium	Symptom-based quick prediction
Support Vector Machine	Classical ML	Small to Medium	80–90%	Medium	Low (with kernel)	Medium	High-dimensional structured data
Decision Tree	Classical ML	Small to Medium	70–80%	Fast	High	High	Easy-to-interpret decision logic
Naïve Bayes	Probabilistic ML	Small to Medium	70–85%	Very Fast	Medium	High	Text or symptom-based data classification
Random Forest	Ensemble (Bagging)	Medium to Large	85–92%	Medium	Low	Medium	General disease prediction
Gradient Boosting	Ensemble (Boosting)	Medium to Large	88–94%	Medium	Medium	Low	Structured/tabular clinical data
XG Boost	Ensemble (Boosting)	Medium to Large	90–96%	Fast	Low	Medium	Top choice for structured datasets
Ada Boost	Ensemble (Boosting)	Small to Medium	80–88%	Medium	Medium	Medium	Noisy or imbalanced datasets
Artificial Neural Network	Deep Learning	Medium to Large	85–92%	Slow	Medium	Low	Capturing complex nonlinear patterns
Convolutional Neural Network	Deep Learning	Large (image data)	92–98%	Slow	Low	Low	Image-based disease diagnosis
Recurrent Neural Network (LSTM)	Deep Learning	Medium to Large (temporal)	85–93%	Slow	Medium	Low	Time-series patient data

Fig 1: Graph showing relation between data set size and algorithm



7. CONCLUSION

This research evaluated twelve machine learning algorithms for disease prediction, revealing that deep learning models like CNN and LSTM achieved the highest accuracy, with CNN reaching up to 94.6% and LSTM close behind at 93.2%. Among traditional algorithms, Random Forest (89.5%) and XGBoost (91.1%) outperformed others like Logistic Regression (82.3%) and KNN (85.6%). Ensemble methods consistently offered improved precision and robustness, while proper data preprocessing significantly boosted model efficiency. The comparative analysis confirms that model selection must align with data characteristics and prediction goals, with deep learning offering superior performance for complex, high-dimensional symptom data.

These findings underline the importance of selecting algorithms based on both dataset structure and prediction complexity. While traditional models offer interpretability and speed, deep learning models excel in handling intricate symptom patterns. Future work should focus on hybrid systems that combine the strengths of multiple algorithms for even greater accuracy and adaptability in real-world healthcare applications.

REFERENCES

1. Khan MN, Srivastava A. Disease prediction using machine learning. Int J Eng Manag Res (IJEMR). 2023 Jun;13(3):23-30.
2. Raju K, Priya H, Supraja M. Multiple disease prediction using machine learning. J Emerg Technol Innov Res (JETIR). 2023 Apr;10(4):88-95.
3. Gaurav K, Kumar A, Singh P, Kumari A, Kasar M, Suryawanshi T. A detailed review on disease prediction models that use machine learning: Human disease prediction using machine learning techniques and real-life parameters. Int J Eng (IJE). 2023 Jun;36(6):78-85.
4. Parshant, Rathee A. Multiple disease prediction using machine learning. IRE J. 2023 Dec;6(12):45-50.

5. hatt A, Singasane S, Chaube N. Disease prediction using machine learning. Int Res J Modern Eng Technol Sci (IRJMETS). 2022 Jan;4(1):11-18.
6. Gomathy CK, Naidu AR. The prediction of disease using machine learning. Int J Sci Res Eng Manag (IJSREM). 2021 Oct;5(10):62-70.
7. Takke K, Bhaijee R, Singh A, Patil A. Medical disease prediction using machine learning algorithms. Int J Res Appl Sci Eng Technol (IJRASET). 2021 May;10(5):102-110.
8. Reddy PP, Babu DM, Kumar H, Sharma S. Disease prediction using machine learning. Int J Creat Res Thoughts (IJCRT). 2021 May;9(5):1234-1242.
9. Farooqui ME, Ahmad J. A detailed review on disease prediction models that use machine learning. Int J Innov Res Comput Sci Technol (IJIRCST). 2020 Jul;8(4):40-48.
10. Chauhan RH, Naik DN, Halpati RA, Patel SJ, Prajapati AD. Disease prediction using machine learning. Int Res J Eng Technol (IRJET). 2020 May;7(5):56-63.