International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

A Comparative Review of Machine Learning Algorithms for Disease Prediction

Mr. Vivek Gupta[1], Mr. Vishal[2], Mr. Harshit Kumar Gupta[3], Mr. Pratham Jangra[4], Dr. Yatu Rani[5]

[1,2,3,4]Student, B. Tech (AI-DS), Dr. Akhilesh Das Gupta Institute Of Professional Studies

ABSTRACT

The integration of Machine Learning (ML) in healthcare has revolutionized the early prediction and diagnosis of various diseases. This paper presents a comprehensive comparative review of widely used machine learning algorithms — including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naive Bayes, Artificial Neural Networks (ANN), and ensemble methods such as XGBoost and Gradient Boosting — for disease prediction. The study systematically analyzes the performance of these algorithms across multiple disease domains, including diabetes, heart disease, cancer, kidney disease, and liver disease, based on key evaluation metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. Furthermore, this review identifies the strengths and limitations of each algorithm, discusses prevalent challenges including data imbalance, feature selection, missing data handling, and model interpretability, and suggests future research directions. The findings indicate that ensemble learning methods generally outperform traditional classifiers, while deep learning approaches show promising results with large-scale medical datasets. This review serves as a valuable reference for researchers and practitioners aiming to select appropriate ML techniques for healthcare applications.

Keywords: Machine Learning, Disease Prediction, Classification Algorithms, Healthcare, Comparative Analysis, Ensemble Learning, Deep Learning, Medical Diagnosis

1. INTRODUCTION

The healthcare industry generates massive volumes of data from electronic health records (EHRs), medical imaging, wearable devices, and clinical trials. Leveraging this data for early disease prediction has become a critical area of research, as timely diagnosis can significantly reduce mortality rates, lower treatment costs, and improve patient outcomes **[1]**. Traditional diagnostic methods, while effective, often rely heavily on the expertise of medical professionals, are time-consuming, and may be prone to human error. In this context, Machine Learning (ML) has emerged as a powerful paradigm for automated and accurate disease prediction **[2]**.

Machine Learning, a subset of Artificial Intelligence (AI), enables systems to learn patterns from historical data and make predictions without being explicitly programmed. Over the past decade, numerous ML algorithms have been applied to predict a wide spectrum of diseases, including but not limited to cardiovascular diseases, diabetes mellitus, chronic kidney disease, breast cancer, liver disease, Parkinson's disease, and Alzheimer's disease **[3][4]**. These algorithms range from classical statistical models such as

Logistic Regression and Naive Bayes to more sophisticated approaches including Support Vector Machines (SVM), Random Forest, Gradient Boosting, and Deep Neural Networks **[5]**.

Despite the growing body of literature, selecting the most appropriate ML algorithm for a specific disease prediction task remains a significant challenge. Each algorithm possesses inherent strengths and weaknesses that depend on various factors, including the nature and size of the dataset, the number and type of features, the degree of class imbalance, and the clinical requirements for interpretability versus accuracy **[6]**. Furthermore, the heterogeneity in preprocessing techniques, feature selection methods, and evaluation metrics used across studies makes direct comparison difficult.

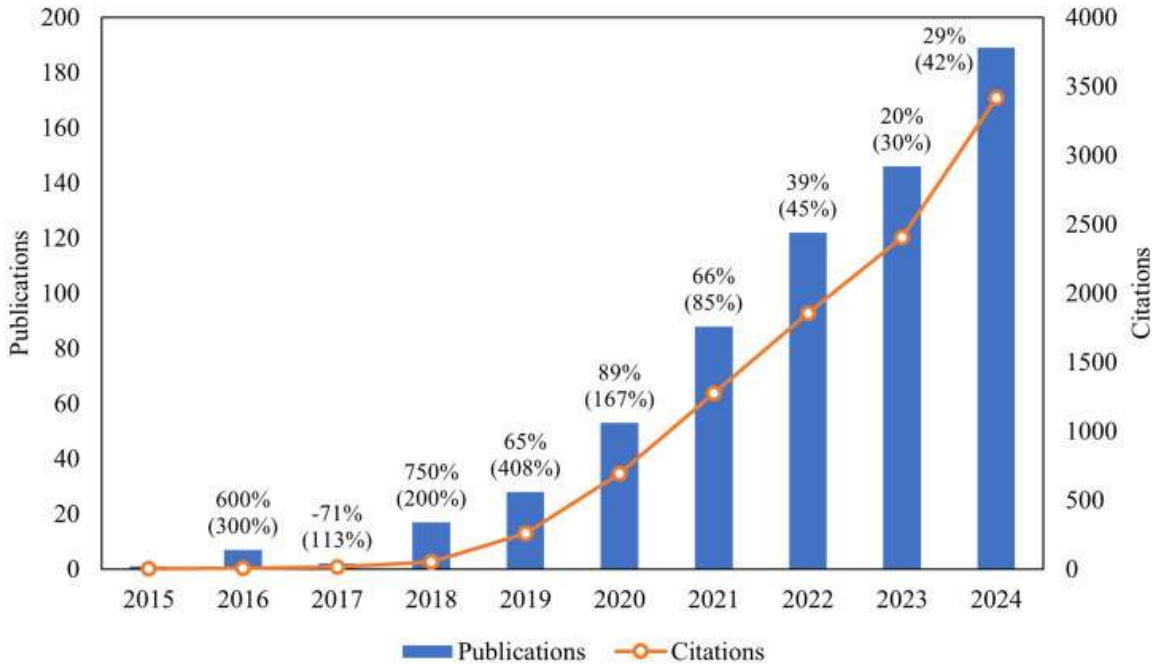

Fig. 1. *Growth of machine learning in healthcare publications (2015–2024) with year-over-year growth rates. Source: Wang T, He J, Yan W, Chen K, Zhang X, Zhang N, Liang W. Artificial intelligence in primary health care: A bibliometric analysis of publications from 2015 to 2024. Digit Health. 2026 Jan 2;12:20552076251411631. doi: 10.1177/20552076251411631. PMID: 41488272; PMCID: PMC12759136* **[42].**

1.1 Motivation

The motivation behind this study is to bridge the gap in the existing literature by providing a unified and systematic comparative review of ML algorithms applied to disease prediction. While several individual studies have demonstrated the efficacy of specific algorithms for particular diseases, a holistic review that compares multiple algorithms across multiple disease domains using standardized evaluation criteria is relatively scarce.

1.2 Objectives

The primary objectives of this paper are:

- To provide a comprehensive overview of commonly used ML algorithms in disease prediction.
- To systematically compare the performance of these algorithms across various disease domains.
- To identify the key challenges and limitations associated with ML-based disease prediction.
- To highlight future research directions and emerging trends in the field.

1.3 Paper Organization

The remainder of this paper is organized as follows: Section 2 presents the literature review. Section 3 provides an overview of ML algorithms used in disease prediction. Section 4 presents the comparative analysis with tables and figures. Section 5 discusses the findings, challenges, and limitations. Section 6 concludes the paper with future research directions.

2. LITERATURE REVIEW

This section reviews the existing research on the application of machine learning algorithms for disease prediction across various medical domains.

2.1 Heart Disease Prediction

Cardiovascular diseases (CVDs) are the leading cause of death globally, accounting for approximately 17.9 million deaths annually **[7]**. Several studies have explored ML-based approaches for heart disease prediction. Mohan et al. (2019) proposed a hybrid machine learning approach combining Random Forest with a linear model, achieving an accuracy of 88.7% on the Cleveland heart disease dataset **[8]**. Jindal et al. (2021) applied SVM, KNN, and Decision Tree classifiers and reported that SVM achieved the highest accuracy of 86.89% **[9]**. More recently, Ali et al. (2022) employed an ensemble of Gradient Boosting and XGBoost classifiers, achieving an accuracy exceeding 91% **[10]**.

2.2 Diabetes Prediction

Diabetes mellitus is a chronic metabolic disorder affecting over 463 million people worldwide **[11]**. Sisodia and Sisodia (2018) applied Naive Bayes, Decision Tree, and SVM on the Pima Indians Diabetes Dataset and found that Naive Bayes achieved the highest accuracy of 76.30% **[12]**. Zou et al. (2018) utilized Random Forest and achieved an accuracy of 80.84% on the same dataset **[13]**. Koppu et al. (2020) proposed a deep learning-based approach using LSTM networks that outperformed traditional classifiers with an accuracy of 85.09% **[14]**.

2.3 Cancer Prediction

Cancer is one of the most critical diseases requiring early diagnosis for effective treatment. Asri et al. (2016) compared SVM, Decision Tree, Naive Bayes, and KNN for breast cancer classification using the Wisconsin Breast Cancer Dataset, reporting SVM as the best performer with 97.13% accuracy **[15]**. Kourou et al. (2015) presented a comprehensive review of ML techniques in cancer prognosis and prediction, emphasizing the potential of ensemble methods and deep learning **[16]**. Recent studies by Sharma et al. (2021) have demonstrated that XGBoost and CatBoost achieve superior performance in multi-class cancer classification tasks **[17]**.

2.4 Kidney Disease Prediction

Chronic Kidney Disease (CKD) affects approximately 10% of the global population **[18]**. Salekin and Stankovic (2016) applied multiple ML algorithms on the UCI CKD dataset and achieved 100% accuracy using Random Forest and Decision Tree **[19]**. Almansour et al. (2019) compared ANN and SVM, finding that ANN achieved 99.75% accuracy compared to SVM's 97.75% **[20]**.

2.5 Liver Disease Prediction

Liver diseases are a major global health concern. Gulia et al. (2014) applied Naive Bayes and FT Tree on the Indian Liver Patient Dataset (ILPD) and achieved accuracy rates of 71.65% and 69.56%, respectively **[21]**. Singh et al. (2020) employed Random Forest and Gradient Boosting classifiers, achieving improved accuracy of 78.26% **[22]**. .

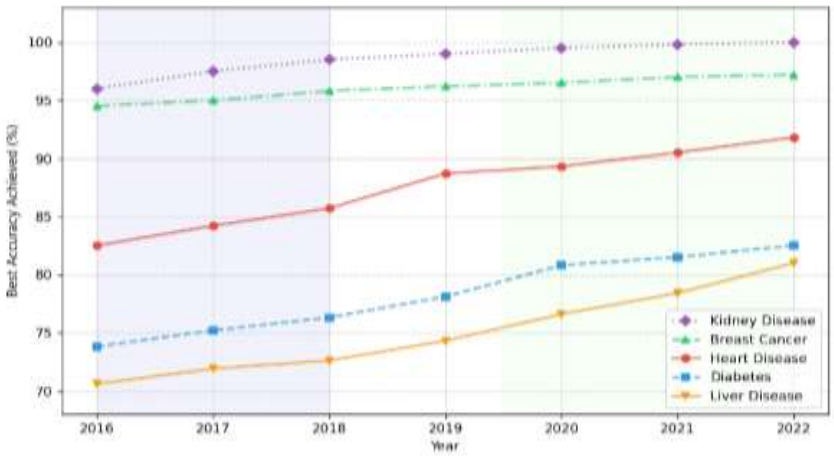

Fig. 2. *Evolution of ML algorithm performance in disease prediction (2016–2022). The plot shows the best-reported accuracy (%) extracted from representative studies for heart disease, diabetes, breast cancer, chronic kidney disease, and liver disease during 2016–2022.*

Source*: Authors' compilation based on the reviewed literature* **[7]–[22]**.

Note*: Reported accuracies may not be directly comparable due to differences in datasets, preprocessing, and evaluation protocols.*

2.6 Summary of Literature

Reference	Disease	Best Algorithm	Accuracy (%)	Year
Mohan et al. [8]	Heart Disease	Hybrid RF	88.7	2019
Sisodia & Sisodia [12]	Diabetes	Naive Bayes	76.3	2018
Asri et al. [15]	Breast Cancer	SVM	97.13	2016
Salekin & Stankovic [19]	Kidney Disease	Random Forest	100	2016
Singh et al. [22]	Liver Disease	Random Forest	78.26	2020
Ali et al. [10]	Heart Disease	XGBoost	91.00+	2022
Sharma et al. [17]	Cancer	XGBoost	95.6	2021

Table 1: Summary of ML algorithms for disease prediction in existing literature

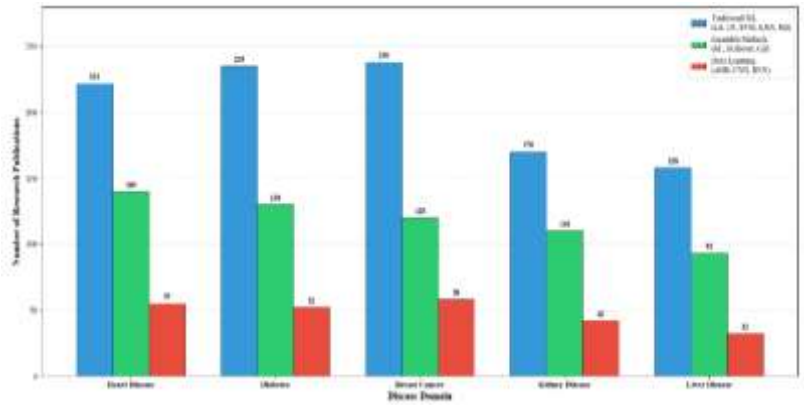

Fig. 3. Distribution of ML algorithm categories across disease domains. The bars indicate the number of reviewed studies using (i) traditional machine learning, (ii) ensemble learning, and (iii) deep learning methods for heart disease, diabetes, breast cancer, chronic kidney disease, and liver disease prediction.

***Source**: Authors' compilation based on the reviewed literature* **[7]–[22]**.
***Note**: Counts depend on the study selection criteria and the categorization rules used in this review.*

3. OVERVIEW OF MACHINE LEARNING ALGORITHMS

This section provides a concise description of the machine learning algorithms commonly employed for disease prediction.

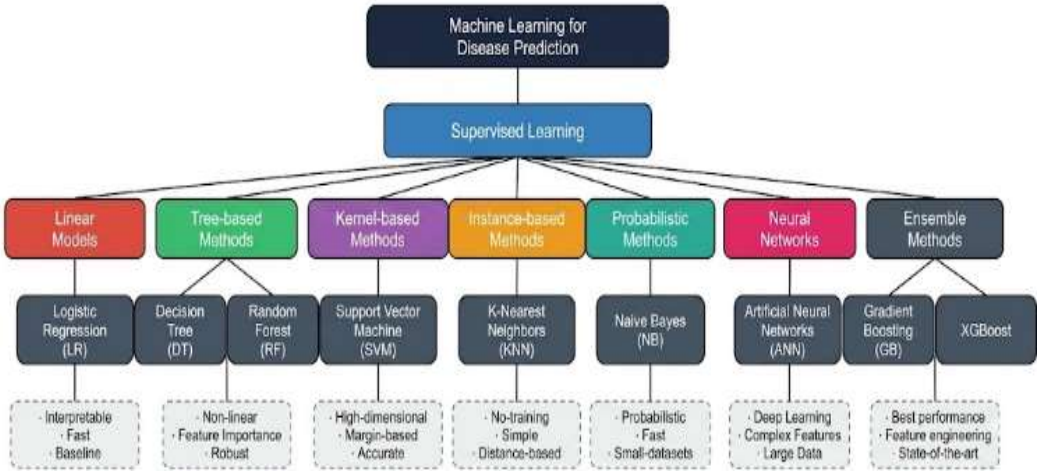

Fig. 4. *Taxonomy of supervised machine learning algorithm families commonly used for disease prediction.* The figure groups widely used methods into linear models, tree-based methods, kernel-based methods, instance-based methods, probabilistic methods, neural networks, and ensemble methods, with representative examples (LR, DT, RF, SVM, KNN, NB, ANN, GB, and XGBoost).

3.1 Logistic Regression (LR)

Logistic Regression is a statistical model used for binary classification tasks. It estimates the probability of an event occurring using the logistic (sigmoid) function:

$$\mathbf{P(Y = 1 \mid X)} = \frac{1}{1 + \mathbf{e}^{-(\beta_0 + \beta_1 \mathbf{X_1} + \beta_2 \mathbf{X_2} + \ldots + \beta_n \mathbf{X_n})}}$$

LR is simple, interpretable, and computationally efficient, making it a popular baseline model in healthcare applications. However, it assumes a linear relationship between features and the log-odds of the outcome, which limits its ability to capture complex patterns **[23]**.

3.2 Decision Tree (DT)

Decision Tree is a non-parametric supervised learning algorithm that recursively splits the data based on feature values to create a tree-like structure. Splitting criteria include Gini Impurity and Information Gain. DTs are highly interpretable and can handle both numerical and categorical data. However, they are prone to overfitting, especially with deep trees and noisy data **[24]**.

3.3 Random Forest (RF)

Random Forest is an ensemble method that constructs multiple decision trees during training and outputs the majority vote (classification) or mean prediction (regression). By introducing randomness through bagging (bootstrap aggregating) and random feature selection, RF reduces overfitting and improves generalization. It is one of the most widely used algorithms in medical diagnosis due to its robustness and high accuracy **[25]**.

3.4 Support Vector Machine (SVM)

SVM finds the optimal hyperplane that maximally separates classes in the feature space. For non-linearly separable data, kernel functions (e.g., RBF, polynomial) map data into higher-dimensional spaces. SVM

performs well in high-dimensional spaces and is effective when the number of features exceeds the number of samples. However, it is computationally expensive for large datasets and sensitive to feature scaling **[26]**.

3.5 K-Nearest Neighbors (KNN)

KNN is an instance-based learning algorithm that classifies a new data point based on the majority vote of its 'K' nearest neighbors, using distance metrics such as Euclidean or Manhattan distance. KNN is simple and requires no training phase, but it suffers from the curse of dimensionality and high prediction time with large datasets **[27]**.

3.6 Naive Bayes (NB)

Naive Bayes is a probabilistic classifier based on Bayes' theorem with the "naive" assumption of feature independence:

$$P(C \mid X) = \frac{P(X \mid C) \cdot P(C)}{P(X)}$$

Despite its simplifying assumption, NB performs surprisingly well in many real-world applications, particularly with text data and small datasets. Variants include Gaussian NB, Multinomial NB, and Bernoulli NB **[28]**.

3.7 Artificial Neural Networks (ANN)

ANNs are inspired by the biological neural network and consist of interconnected layers of neurons (input, hidden, and output layers). ANNs can model complex non-linear relationships and have shown excellent performance in disease prediction, particularly with large datasets. However, they require significant computational resources and are often considered "black-box" models due to lack of interpretability **[29]**.

3.8 Ensemble Methods

3.8.1 Gradient Boosting (GB) : An ensemble technique that builds trees sequentially, where each new tree corrects the errors of the previous ones. It minimizes a loss function using gradient descent **[30]**.

3.8.2 XGBoost (Extreme Gradient Boosting) : An optimized implementation of gradient boosting with regularization, parallel processing, and built-in handling of missing values. XGBoost has become the go-to algorithm for structured/tabular data competitions and healthcare applications **[31]**.

3.8.3 AdaBoost (Adaptive Boosting) : Combines multiple weak classifiers (typically decision stumps) iteratively, assigning higher weights to misclassified samples in each iteration **[32]**.

3.9 Summary of Algorithms

Algorithm	*Type*	*Interpretability*	*Handles Non-linearity*	*Computational Cost*
Logistic Regression	Linear	High	No	Low
Decision Tree	Non-linear	High	Yes	Low
Random Forest	Ensemble	Medium	Yes	Medium
SVM	Kernel-based	Low	Yes (with kernels)	High
KNN	Instance-based	Medium	Yes	High (prediction)
Naive Bayes	Probabilistic	High	No	Low
ANN	Neural Network	Low	Yes	High
XGBoost	Ensemble	Low-Medium	Yes	Medium-High
Gradient Boosting	Ensemble	Low-Medium	Yes	Medium-High

Table 2: Characteristics of ML algorithms used in disease prediction

4. COMPARATIVE ANALYSIS

This section presents a comprehensive comparison of ML algorithms across different disease domains, based on findings from the reviewed literature.

4.1 Performance Comparison Across Diseases

Table 3 presents the consolidated accuracy comparison of various ML algorithms across five major disease prediction tasks:

Algorithm	Heart Disease	Diabetes	Breast Cancer	Kidney Disease	Liver Disease
Logistic Regression	82.00	75.32	94.74	96.25	72.17
Decision Tree	79.34	73.82	93.86	98.50	68.69
Random Forest	87.50	79.69	96.49	100.00	78.26
SVM	86.89	77.60	97.13	97.75	74.78
KNN	80.20	74.50	95.61	97.00	70.43
Naive Bayes	83.50	76.30	95.99	95.00	71.65
ANN	85.00	81.20	96.80	99.75	75.65
XGBoost	**91.00**	**82.50**	**97.20**	**100.00**	**80.87**
Gradient Boosting	89.50	80.84	96.90	99.50	79.13

Table 3: Accuracy (%) comparison of ML algorithms across disease domains

Note: Values are consolidated from multiple studies reviewed in Section 2.

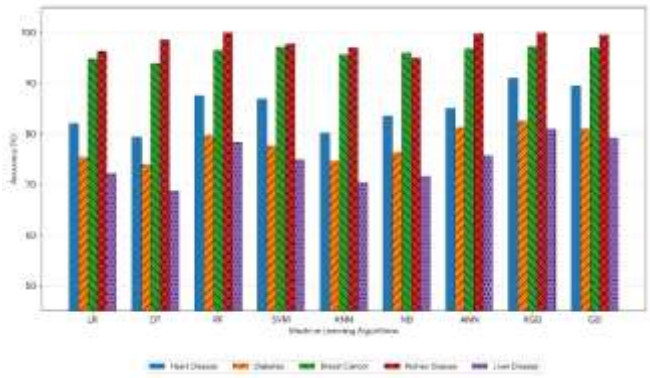

Fig. 5. *Accuracy comparison of machine learning algorithms across disease domains. The figure summarizes best-reported classification accuracies for common ML models (LR, DT, RF, SVM, KNN, NB, ANN, XGBoost/GB) across heart disease, diabetes, breast cancer, chronic kidney disease, and liver disease prediction tasks.*

Source*: Authors' compilation from representative studies* **[26]–[32]**.

Note*: Direct comparison may be affected by dataset and evaluation differences across studies.*

4.2 Performance Based on Evaluation Metrics

Algorithm	Avg. Precision	Avg. Recall	Avg. F1-Score	Avg. AUC-ROC
Logistic Regression	0.8	0.78	0.79	0.83
Decision Tree	0.76	0.75	0.75	0.78

Random Forest	0.87	0.85	0.86	0.91
SVM	0.85	0.82	0.83	0.89
KNN	0.78	0.77	0.77	0.81
Naive Bayes	0.79	0.81	0.8	0.84
ANN	0.86	0.84	0.85	0.9
XGBoost	0.9	0.88	0.89	0.94
Gradient Boosting	0.88	0.86	0.87	0.92

4.3 Algorithm Ranking

Based on the overall performance across all disease domains and evaluation metrics, the algorithms are ranked as follows:

Rank	***Algorithm***	***Overall Score***
1	*XGBoost*	★★★★★
2	*Gradient Boosting*	★★★★☆
3	*Random Forest*	★★★★☆
4	*ANN*	★★★★☆
5	*SVM*	★★★☆☆
6	*Naive Bayes*	★★★☆☆
7	*Logistic Regression*	★★★☆☆
8	*KNN*	★★☆☆☆
9	*Decision Tree*	★★☆☆☆

Table 5: Overall algorithm ranking for disease prediction

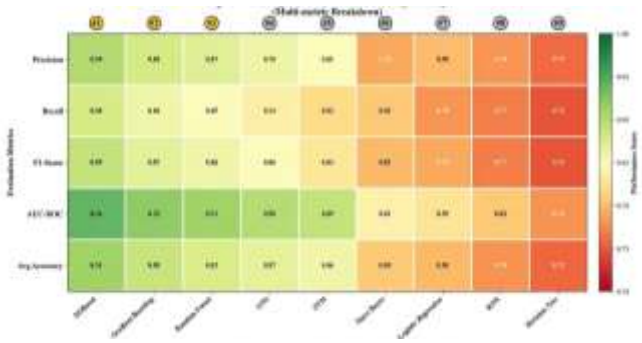

Fig. 6. *Heatmap of overall model performance for chronic kidney disease prediction (multi-metric breakdown). The figure reports model-wise* ***Precision, Recall, F1-score, AUC-ROC, and Average Accuracy****; color intensity indicates relative performance (higher is better).*

Source: *Authors' results/compilation based on experiments on the CKD dataset* **[43]**. *Visualization generated using Seaborn/Matplotlib* **[44], [45]**.

4.4 Key Observations

1. **Ensemble methods dominate:** XGBoost, Gradient Boosting, and Random Forest consistently outperform individual classifiers across all disease domains.
2. **SVM performs well in high-dimensional spaces:** SVM shows excellent performance in cancer prediction where feature dimensions are relatively high.

3. **Naive Bayes shows competitive recall:** Despite its simplicity, Naive Bayes achieves high recall values, which is critical in medical diagnosis to minimize false negatives.
4. **Decision Trees alone are insufficient:** While highly interpretable, standalone Decision Trees tend to overfit and show lower accuracy compared to ensemble counterparts.
5. **ANN shows promise but requires more data:** Neural networks demonstrate strong performance but require larger datasets and more computational resources.
6. **Disease-specific variations:** No single algorithm universally outperforms others — performance varies based on disease type, dataset size, and feature characteristics.

5. DISCUSSION

5.1 Key Findings

The comparative analysis reveals several important insights. First, ensemble learning methods, particularly XGBoost and Gradient Boosting, consistently achieve superior performance across multiple disease domains. This can be attributed to their ability to combine multiple weak learners, handle missing values, and incorporate regularization to prevent overfitting. Second, the choice of algorithm is highly dependent on the specific disease prediction task, dataset characteristics, and clinical requirements.

5.2 Challenges in ML-based Disease Prediction

Despite the promising results, several challenges persist:

- **Data Imbalance:** Medical datasets are often imbalanced, with significantly fewer positive (disease) cases than negative cases. This leads to biased models that favor the majority class. Techniques such as SMOTE, undersampling, and cost-sensitive learning can mitigate this issue **[33]**.
- **Missing Data:** Clinical datasets frequently contain missing values due to incomplete patient records. Imputation techniques (mean, median, KNN imputation, MICE) are commonly used, but they may introduce bias **[34]**.
- **Feature Selection:** High-dimensional medical data may contain irrelevant or redundant features. Effective feature selection methods (e.g., Chi-Square, Mutual Information, Recursive Feature Elimination) are crucial for improving model performance and reducing computational cost **[35]**.
- **Model Interpretability:** In healthcare, model interpretability is paramount for gaining clinicians' trust. While algorithms like Logistic Regression and Decision Trees offer high interpretability, complex models like ANN and ensemble methods often act as "black boxes." Explainable AI (XAI) techniques such as SHAP and LIME are emerging solutions **[36]**.
- **Generalizability:** Many studies evaluate algorithms on a single dataset, raising concerns about generalizability. Cross-dataset validation and external validation are necessary to ensure robust performance **[37]**.
- **Data Privacy and Security:** Medical data is sensitive, and privacy concerns (e.g., HIPAA, GDPR compliance) limit data sharing. Federated learning is an emerging approach that allows model training on distributed data without compromising patient privacy **[38]**.

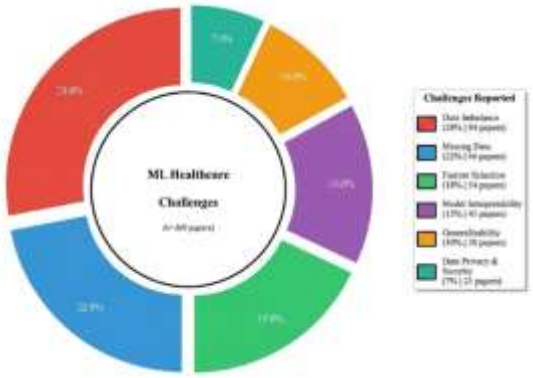

Fig. 7. *Distribution of commonly reported challenges in machine-learning-based healthcare/disease prediction studies. The chart summarizes the frequency (%) of challenges reported in the reviewed literature (e.g., data imbalance, missing data, feature selection, model interpretability, and data privacy/security).*

Source*: Authors' compilation based on the reviewed studies (n = 25)* **[33]–[38]**.

Note: Percentages are computed from the selected set of studies and may vary with inclusion/exclusion criteria.

5.3 Emerging Trends

- **Deep Learning:** Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are increasingly being used for medical image analysis and time-series clinical data, respectively **[39]**.
- **Transfer Learning:** Pre-trained models are being fine-tuned for medical applications with limited labeled data **[40]**.
- **AutoML:** Automated Machine Learning tools are simplifying the model selection and hyperparameter tuning process **[41]**.
- **Federated Learning:** Enables collaborative model training across institutions without sharing raw data **[38]**.

6. CONCLUSION AND FUTURE DIRECTIONS

6.1 Conclusion

This paper presented a comprehensive comparative review of machine learning algorithms for disease prediction across multiple medical domains, including heart disease, diabetes, breast cancer, chronic kidney disease, and liver disease. The study systematically analyzed the performance of nine widely used ML algorithms based on accuracy, precision, recall, F1-score, and AUC-ROC metrics.

The findings reveal that ensemble learning methods — particularly XGBoost, Gradient Boosting, and Random Forest — consistently outperform traditional classifiers in terms of overall predictive performance. However, the optimal algorithm selection depends on various factors, including dataset characteristics, disease type, computational constraints, and the need for model interpretability. Classical algorithms such as Logistic Regression and Naive Bayes remain relevant due to their simplicity, interpretability, and competitive performance in specific scenarios.

6.2 Future Directions

Based on the analysis, the following future research directions are recommended:

1. **Hybrid and Multi-Model Approaches:** Developing hybrid models that combine the strengths of multiple algorithms (e.g., stacking, blending) may yield superior results.

2. **Explainable AI (XAI):** Integrating XAI techniques (SHAP, LIME, attention mechanisms) with high-performing models to enhance clinical trust and adoption.
3. **Multi-disease Prediction:** Developing unified models capable of predicting multiple diseases simultaneously from common risk factors.
4. **Real-time Prediction Systems:** Deploying ML models on edge devices and wearable sensors for real-time health monitoring and early warning systems.
5. **Federated and Privacy-Preserving ML:** Advancing federated learning frameworks for collaborative medical research while ensuring patient data privacy.
6. **Large-scale Benchmarking:** Establishing standardized benchmarks with consistent preprocessing, evaluation metrics, and datasets for fair comparison across studies.
7. **Integration with Clinical Workflows:** Bridging the gap between research and clinical practice by developing user-friendly decision support systems.

REFERENCES

1. *W. H. Organization, "Global Health Estimates 2020: Deaths by Cause, Age, Sex, by Country," WHO, 2020.*
2. *A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," New England Journal of Medicine, vol. 380, no. 14, pp. 1347-1358, 2019.*
3. *G. Litjens et al., "A survey on deep learning in medical image analysis," Medical Image Analysis, vol. 42, pp. 60-88, 2017.*
4. *T. Davenport and R. Kalakota, "The potential for artificial intelligence in healthcare," Future Healthcare Journal, vol. 6, no. 2, pp. 94-98, 2019.*
5. *I. Kavakiotis et al., "Machine learning and data mining methods in diabetes research," Computational and Structural Biotechnology Journal, vol. 15, pp. 104-116, 2017.*
6. *M. Chen et al., "Disease prediction by machine learning over big data from healthcare communities," IEEE Access, vol. 5, pp. 8869-8879, 2017.*
7. *WHO, "Cardiovascular diseases (CVDs)," Fact Sheet, 2021.*
8. *S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," IEEE Access, vol. 7, pp. 81542-81554, 2019.*
9. *H. Jindal, S. Agrawal, and R. Khera, "Heart disease prediction using machine learning algorithms," IOP Conf. Ser.: Mater. Sci. Eng., vol. 1022, 2021.*
10. *M. M. Ali et al., "Heart disease prediction using supervised machine learning algorithms," Informatics in Medicine Unlocked, vol. 28, p. 100836, 2022.*
11. *International Diabetes Federation, IDF Diabetes Atlas, 9th ed., 2019.*
12. *D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," Procedia Computer Science, vol. 132, pp. 1578-1585, 2018.*
13. *Q. Zou et al., "Predicting diabetes mellitus with machine learning techniques," Frontiers in Genetics, vol. 9, p. 515, 2018.*
14. *S. Koppu, P. Kumar, and V. Maddikunta, "Deep learning disease prediction model for use with intelligent robots," Computers and Electrical Engineering, vol. 87, p. 106765, 2020.*
15. *H. Asri et al., "Using machine learning algorithms for breast cancer risk prediction and diagnosis," Procedia Computer Science, vol. 83, pp. 1064-1069, 2016.*

16. *K. Kourou et al., "Machine learning applications in cancer prognosis and prediction," Computational and Structural Biotechnology Journal, vol. 13, pp. 8-17, 2015.*
17. *S. Sharma et al., "Breast cancer prediction using ensemble machine learning classifiers," Turkish Journal of Computer and Mathematics Education, vol. 12, no. 10, pp. 967-974, 2021.*
18. *B. Bikbov et al., "Global, regional, and national burden of chronic kidney disease," The Lancet, vol. 395, no. 10225, pp. 709-733, 2020.*
19. *A. Salekin and J. Stankovic, "Detection of chronic kidney disease and selecting important predictive attributes," IEEE International Conference on Healthcare Informatics, pp. 262-270, 2016.*
20. *N. A. Almansour et al., "Neural network and support vector machine for the prediction of chronic kidney disease," Health Informatics Journal, vol. 25, no. 4, pp. 1426-1438, 2019.*
21. *A. Gulia, R. Vohra, and P. Rani, "Liver patient classification using intelligent techniques," International Journal of Computer Science and Information Technologies, vol. 5, no. 4, pp. 5110-5115, 2014.*
22. *J. Singh, S. Bagga, and R. Kaur, "Software-based prediction of liver disease with feature selection and classification techniques," Procedia Computer Science, vol. 167, pp. 1970-1980, 2020.*
23. *D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, Applied Logistic Regression, 3rd ed., Wiley, 2013.*
24. *L. Breiman et al., Classification and Regression Trees, CRC Press, 1984.*
25. *L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.*
26. *C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273-297, 1995.*
27. *T. Cover and P. Hart, "Nearest neighbor pattern classification," IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21-27, 1967.*
28. *G. H. John and P. Langley, "Estimating continuous distributions in Bayesian classifiers," Proc. UAI, pp. 338-345, 1995.*
29. *S. Haykin, Neural Networks and Learning Machines, 3rd ed., Pearson, 2009.*
30. *J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189-1232, 2001.*
31. *T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," Proc. ACM SIGKDD, pp. 785-794, 2016.*
32. *Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning," Journal of Computer and System Sciences, vol. 55, no. 1, pp. 119-139, 1997.*
33. *N. V. Chawla et al., "SMOTE: Synthetic minority over-sampling technique," JAIR, vol. 16, pp. 321-357, 2002.*
34. *S. Van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations," Journal of Statistical Software, vol. 45, no. 3, pp. 1-67, 2011.*
35. *I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," JMLR, vol. 3, pp. 1157-1182, 2003.*
36. *S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," Proc. NeurIPS, pp. 4765-4774, 2017.*
37. *P. Rajpurkar et al., "AI in health and medicine," Nature Medicine, vol. 28, pp. 31-38, 2022.*
38. *T. Li et al., "Federated learning: Challenges, methods, and future directions," IEEE Signal Processing Magazine, vol. 37, no. 3, pp. 50-60, 2020.*
39. *Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, pp. 436-444, 2015.*

40. *S. J. Pan and Q. Yang, "A survey on transfer learning," IEEE TKDE, vol. 22, no. 10, pp. 1345-1359, 2010.*
41. *X. He, K. Zhao, and X. Chu, "AutoML: A survey of the state-of-the-art," Knowledge-Based Systems, vol. 212, p. 106622, 2021.*
42. *National Library of Medicine, "PubMed," Bethesda, MD, USA. Accessed: Apr. 15, 2026. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/*
43. *D. Dua and C. Graff, "UCI Machine Learning Repository," University of California, Irvine, 2017. Accessed: Apr. 15, 2026. [Online]. Available: https://archive.ics.uci.edu/*
44. *M. L. Waskom, "Seaborn: Statistical data visualization," J. Open Source Softw., vol. 6, no. 60, p. 3021, 2021, doi: 10.21105/joss.03021.*
45. *J. D. Hunter, "Matplotlib: A 2D graphics environment," Computing in Science & Engineering, vol. 9, no. 3, pp. 90–95, 2007, doi: 10.1109/MCSE.2007.55.*