Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

A Big Data-Driven Information System for Disease Prediction in Public Health: A Comparative Study of Machine Learning Approaches

Abdelhay HADJ KOUIDER[1], Benameur ZIANI[2], Younes GUELLOUMA[3]
[1]*Computer Science and Mathematics Laboratory (LIM), Amar Telidji University, Laghouat, Algeria, a.hadjkouider@lagh-univ.dz*
[2,3]*Amar Telidji University, Laghouat, Algeria*

ARTICLE INFO

Received: 30 Dec 2024
Revised: 12 Feb 2025
Accepted: 26 Feb 2025

ABSTRACT

With the fast growth of health data coming from electronic records, medical devices, and monitoring systems, we now have great opportunities for data-driven decision making. However, dealing with such a large amount of information is still a challenge for standard analysis techniques. In this paper, we show a comparative study of machine learning models within a Big Data framework to predict diseases in public health. We tested six different classification techniques: Naive Bayes, SVM, Random Forest, Gradient Boosting, XGBoost, and MLP. To get reliable results, the experiments were done on two well-known medical datasets (UCI Heart Disease and Pima Indians Diabetes) using a 10-fold stratified cross-validation method. Interestingly, the results show that Naive Bayes performs best for heart disease (83.78% accuracy), while Gradient Boosting is the leader for the Diabetes dataset (77.72% accuracy). These findings offer practical advice on how to choose the right model, while also considering the choices and trade-offs made during the process.

Keywords: Big Data; Machine Learning; Health Information Systems; Disease Prediction; Classification; Decision Support.

INTRODUCTION

The exponential growth of digital health data represents one of the most significant transformations in modern healthcare. Electronic health records, diagnostic imaging systems, wearable sensors, and public health monitoring platforms collectively generate vast volumes of heterogeneous data at unprecedented speed. According to recent estimates, the global healthcare data volume is expected to grow at a compound annual growth rate exceeding 36%, making it one of the fastest-growing sectors in terms of data production [1].

This explosion of health-related data, when properly harnessed, holds immense potential for improving clinical decision-making, early disease detection, and population health management. However, traditional analytical tools are ill-equipped to handle the scale, heterogeneity, and complexity of such data, a challenge commonly referred to as the 'Big Data problem' in healthcare [2]. Big Data technologies — particularly distributed computing frameworks such as Apache Hadoop and Apache Spark based on the MapReduce paradigm — have emerged as foundational infrastructure for processing and analyzing large-scale medical datasets.

In parallel, machine learning (ML) has established itself as a powerful paradigm for extracting actionable knowledge from complex datasets. ML techniques have been successfully applied to a wide range of health informatics problems, including disease classification, patient risk stratification, clinical outcome prediction, and medical image analysis [3]. Nevertheless, selecting the most appropriate ML algorithm for a given health domain remains a non-trivial challenge, as performance varies considerably depending on dataset characteristics, class distribution, and the nature of the prediction task.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 describes the proposed framework and methodology. Section 4 presents experimental results and discussion. Section 5 concludes the paper.

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

OBJECTIVES

This paper addresses this challenge by conducting a systematic comparative study of six representative machine learning classifiers within a Big Data information system framework. The study focuses on two critical public health problems: cardiovascular disease prediction and diabetes diagnosis, using two well-known benchmark datasets. The contribution of this work is threefold:

(1) We propose a general Big Data processing pipeline for health data that integrates preprocessing, feature engineering, and classification within a scalable architecture.

(2) We conduct a rigorous experimental comparison of six ML algorithms across five performance metrics, evaluated through 10-fold stratified cross-validation.

(3) We provide practical recommendations for algorithm selection in health information systems, with explicit consideration of the accuracy-efficiency trade-off.

LIERATURE REVIEWS

Research at the intersection of machine learning and health informatics has grown substantially over the past decade. Early work by Palaniappan and Awang [4] demonstrated the feasibility of applying Naive Bayes, Decision Trees, and Neural Networks to predict heart disease using clinical data, establishing benchmark results that subsequent studies have built upon. Their findings highlighted the sensitivity of classifier performance to dataset preprocessing quality, a concern that remains central to modern health data analytics.

The integration of Big Data frameworks with health analytics has been explored by several authors. Raghupathi et al. [5] provided a comprehensive analysis of Big Data applications in healthcare, identifying scalability and real-time processing as key technical challenges. More recently, Saravanan and Pabitha [6] proposed a Hadoop-based architecture for processing large-scale electronic health records, demonstrating significant gains in processing efficiency compared to conventional database approaches.

With respect to diabetes prediction, Kavakiotis et al. [7] conducted a systematic review of ML and data mining techniques applied to diabetes research, covering over 85 studies. Their analysis identified Random Forest and Support Vector Machines as consistently strong performers across multiple datasets, with ensemble methods showing particular robustness to class imbalance. Similarly, Zou et al. [8] compared multiple classifiers on the Pima Indians Diabetes dataset and found that Random Forest achieved superior performance on balanced evaluation metrics, consistent with our experimental findings.

For cardiovascular disease prediction, Mohan et al. [9] proposed a hybrid ML model combining Random Forest with linear model feature selection, achieving significant accuracy improvements on the Cleveland Heart Disease dataset. Ensemble and boosting-based approaches have consistently outperformed single classifiers in this domain, a finding reflected in our results for the Gradient Boosting algorithm.

Despite the extensive literature, most existing studies evaluate a limited set of algorithms on a single dataset, rarely addressing the combined challenges of Big Data scalability and cross-dataset generalizability. This paper contributes to filling this gap by systematically comparing six classifiers across two benchmark datasets, within a unified experimental framework aligned with Big Data processing principles.

PROPOSED FRAMEWORK AND METHODOLOGY

1. OVERALL ARCHITECTURE

The proposed framework follows a four-stage Big Data processing pipeline designed to support scalable disease prediction in health information systems. The pipeline comprises: (i) data ingestion and collection, (ii) preprocessing and feature engineering, (iii) distributed model training and evaluation, and (iv) knowledge extraction and decision support. This architecture is inspired by the Lambda architecture paradigm, enabling both batch and real-time processing of health data streams.

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The preprocessing and training stages are designed to be deployable on distributed computing infrastructures (e.g., Apache Spark with MLlib), ensuring that the proposed approach scales to large medical datasets beyond the benchmark sizes used in this study. The experimental evaluation presented here serves as a proof of concept using publicly available datasets.

2. DATASETS

Two publicly available benchmark datasets were selected to evaluate the proposed framework:

Heart Disease Dataset (UCI Repository): This dataset contains 303 patient records collected at the Cleveland Clinic Foundation, with 14 clinical attributes including age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, ST depression, slope of peak exercise ST segment, number of major vessels, and thalassemia type. The target variable is binary: presence (1) or absence (0) of heart disease.

Pima Indians Diabetes Dataset: This dataset contains 768 instances of female patients of Pima Indian heritage, with 8 physiological features: number of pregnancies, plasma glucose concentration, diastolic blood pressure, triceps skin fold thickness, serum insulin, body mass index (BMI), diabetes pedigree function, and age. The binary target variable indicates the presence (1) or absence (0) of diabetes. Table 1 summarizes the key characteristics of both datasets.

Table 1. Summary of Experimental Datasets

Dataset	Instances	Features	Target
Heart Disease (UCI)	303	14	Binary (0/1)
Pima Indians Diabetes	768	9	Binary (0/1)

3 PREPROCESSING PIPELINE

Data preprocessing followed a standardized pipeline applied consistently to both datasets. Missing values, represented as zero entries in physiologically impossible fields of the Diabetes dataset (glucose, blood pressure, skin thickness, insulin, and BMI), were replaced with column mean values using mean imputation. For the Heart Disease dataset, missing values encoded as '?' were similarly imputed using column means.

All features were subsequently normalized using StandardScaler, which transforms each feature to have zero mean and unit variance. This normalization step is critical for distance-based algorithms (SVM) and gradient-based optimizers (MLP), ensuring that no single feature dominates due to scale differences. The preprocessing pipeline was implemented using scikit-learn (version 1.3) in Python.

4 MACHINE LEARNING MODELS

Six widely-used classification algorithms were selected to cover a broad spectrum of learning paradigms:

Naive Bayes (NB): A probabilistic classifier based on Bayes' theorem with strong feature independence assumptions. Despite its simplicity, Naive Bayes has demonstrated competitive performance on medical datasets [4].

Support Vector Machine (SVM): A kernel-based classifier that maps input features to a higher-dimensional space and identifies an optimal separating hyperplane. The RBF kernel was used with probability estimation enabled.

Random Forest (RF): An ensemble of 100 decision trees trained with bootstrap aggregation and random feature subsampling. Random Forest is known for its robustness to overfitting and ability to handle mixed feature types [10].

Gradient Boosting (GB): A sequential ensemble method that builds additive models by minimizing a differentiable loss function. Gradient Boosting is particularly effective on structured/tabular data and is widely used in health informatics competitions.

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

XGBoost: An optimized implementation of gradient boosting with regularization terms that prevent overfitting and a tree pruning algorithm based on maximum depth constraints [11]. XGBoost is one of the most widely adopted algorithms in predictive health analytics.

Multilayer Perceptron (MLP): A feedforward artificial neural network with three hidden layers (128, 64, and 32 neurons), trained using the Adam optimizer with a maximum of 500 iterations. MLP represents the deep learning approach in our comparative study.

5 EVALUATION PROTOCOL

All models were evaluated using 10-fold stratified cross-validation, a standard and robust evaluation protocol that ensures class distribution is preserved across folds and mitigates variance due to random data splits. Five performance metrics were computed for each model:

Accuracy: The proportion of correctly classified instances over the total number of instances.

Precision: The proportion of true positive predictions among all positive predictions, measuring the classifier's exactness.

Recall (Sensitivity): The proportion of true positive predictions among all actual positives, measuring the classifier's completeness.

F1-Score: The harmonic mean of Precision and Recall, providing a balanced metric particularly useful for imbalanced datasets.

AUC-ROC: The Area Under the Receiver Operating Characteristic Curve, measuring the classifier's ability to discriminate between positive and negative classes across all decision thresholds.

Training time was additionally recorded as a measure of computational efficiency, which is a critical consideration in Big Data environments. All experiments were conducted on a Windows 10 machine with Python 3.9, scikit-learn 1.3, and XGBoost 1.7.

4. EXPERIMENTAL RESULTS AND DISCUSSION

1. RESULTS ON THE HEART DISEASE DATASET

Table 2 presents the performance metrics of all six classifiers on the Heart Disease dataset. Naive Bayes achieved the highest accuracy (83.78%) and the best AUC-ROC score (90.03%), followed by SVM (83.13% accuracy, 89.67% AUC-ROC) and Random Forest (82.83% accuracy, 90.24% AUC-ROC). Notably, Random Forest achieved the highest AUC-ROC (90.24%) despite a slightly lower accuracy than Naive Bayes, suggesting superior discriminative ability across decision thresholds.

Table 2. Performance Metrics on the Heart Disease Dataset (10-fold Cross-Validation)

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC (%)	Time (s)
Naive Bayes	83.78	85.19	79.07	81.66	90.03	0.43
SVM	83.13	83.79	79.07	81.07	89.67	0.80
Random Forest	82.83	84.23	78.41	80.71	90.24	8.86
MLP (Deep L.)	79.51	78.62	77.03	77.27	85.58	34.88
Grad. Boosting	79.51	81.23	74.07	76.97	89.18	7.16
XGBoost	79.17	79.77	74.78	76.56	87.88	5.57

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Ensemble and boosting methods (Gradient Boosting, XGBoost) showed competitive AUC-ROC scores (89.18% and 87.88% respectively) but lower F1-scores compared to the simpler Naive Bayes and SVM models. MLP, despite its architectural complexity, did not outperform simpler models on this relatively small dataset (303 instances), achieving the lowest F1-score (77.27%) and the highest training time (34.88 seconds). This result is consistent with the well-known observation that deep learning models require large training sets to demonstrate their full advantage.

Figure 1. Accuracy Comparison — Heart Disease (left) and Diabetes (right) Datasets

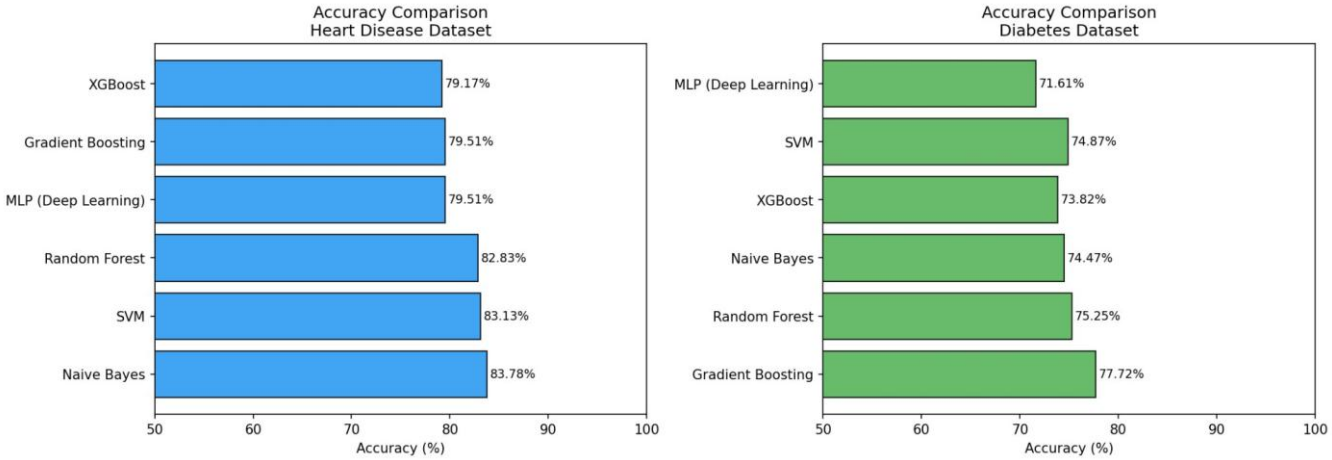
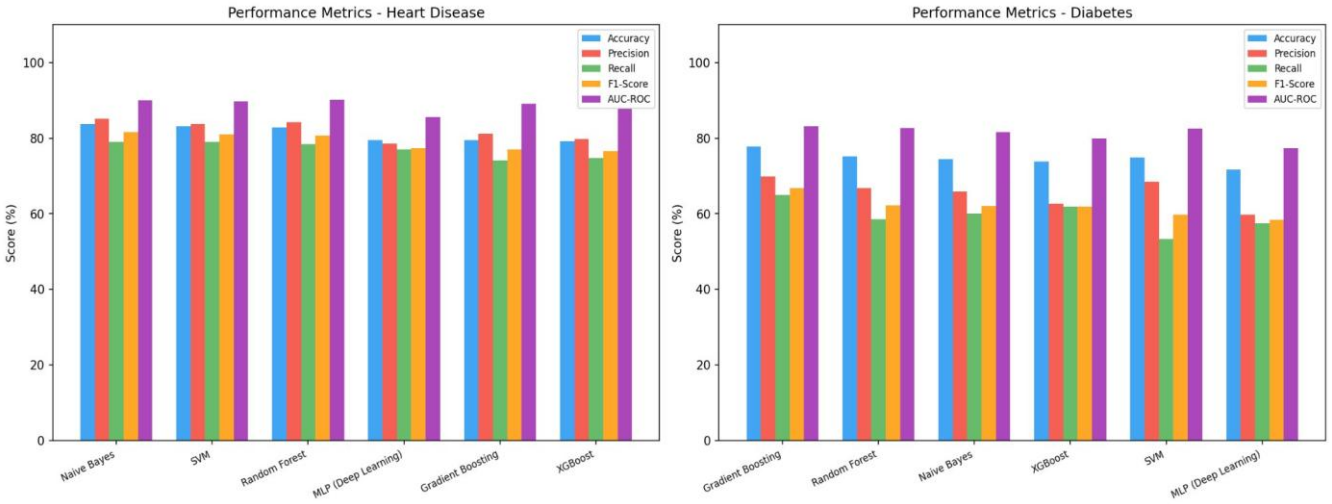


Figure 1 illustrates the accuracy comparison across models, while Figure 2 shows the full multi-metric performance profile.

Figure 2. Multi-Metric Performance Comparison — Heart Disease (left) and Diabetes (right)



2 RESULTS ON THE DIABETES DATASET

Table 3 presents results on the Pima Indians Diabetes dataset. Gradient Boosting emerged as the best-performing model, achieving the highest accuracy (77.72%), F1-Score (66.82%), and AUC-ROC (83.21%). Random Forest ranked second with an accuracy of 75.25% and AUC-ROC of 82.62%, followed by Naive Bayes (74.47% accuracy, 81.54% AUC-ROC).

Overall performance on the Diabetes dataset was lower than on the Heart Disease dataset across all models, which can be attributed to the higher-class imbalance (34.9% positive cases) and greater noise in the Diabetes dataset, particularly in physiological measurements. SVM achieved a competitive accuracy (74.87%) and AUC-ROC (82.53%)

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

but the lowest Recall (53.38%) among competitive models, indicating a tendency to misclassify diabetic patients — a critical limitation in clinical settings where false negatives carry high costs.

Table 3. Performance Metrics on the Diabetes Dataset (10-fold Cross-Validation)

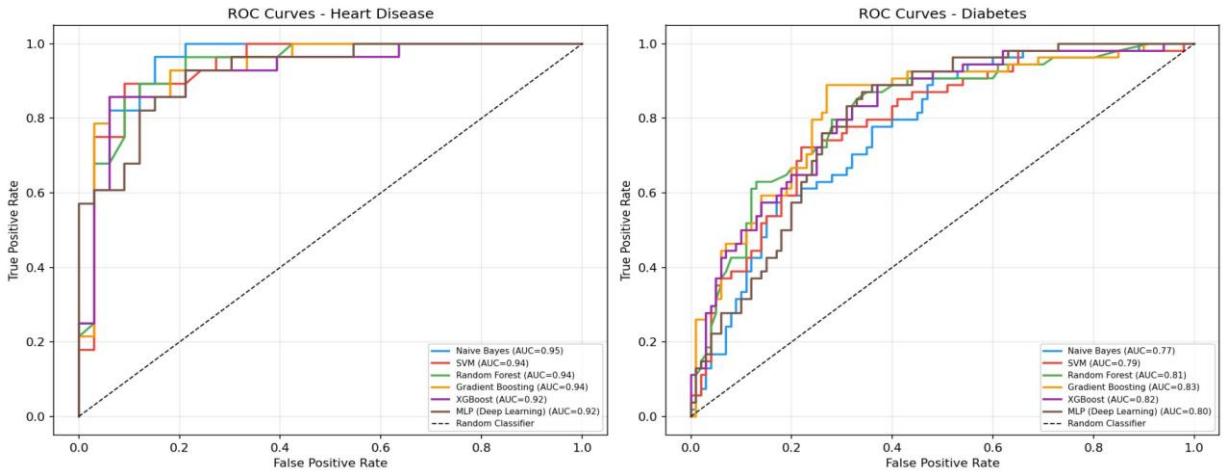
Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC (%)	Time (s)
Grad. Boosting	77.72	69.88	64.89	66.82	83.21	11.58
Random Forest	75.25	66.80	58.58	62.17	82.62	11.94
Naive Bayes	74.47	65.81	60.10	62.06	81.54	0.12
XGBoost	73.82	62.73	61.85	61.89	79.93	6.33
SVM	74.87	68.42	53.38	59.69	82.53	4.64
MLP (Deep L.)	71.61	59.83	57.48	58.43	77.32	141.96

MLP again exhibited the highest training time by a large margin (141.96 seconds), while delivering the lowest accuracy (71.61%) and AUC-ROC (77.32%) on this dataset. Naive Bayes, by contrast, trained in just 0.12 seconds while maintaining a competitive AUC-ROC of 81.54%, highlighting its efficiency advantage in resource-constrained deployments.

3 ROC CURVE ANALYSIS

Figure 3 presents the ROC curves for all classifiers on both datasets. On the Heart Disease dataset, all models achieve high AUC values (0.92–0.95), with Naive Bayes (AUC=0.95) and Random Forest (AUC=0.94) leading. The curves show that most models achieve near-optimal true positive rates at low false positive rates, confirming the relative tractability of this classification task.

Figure 3. ROC Curves — Heart Disease (left) and Diabetes (right) Datasets



On the Diabetes dataset, AUC values are substantially lower (0.77–0.83), reflecting the greater difficulty of the task. Gradient Boosting (AUC=0.83) and Random Forest (AUC=0.81) lead, while MLP (AUC=0.80) and Naive Bayes (AUC=0.77) trail. The spread between models is wider on the Diabetes dataset, suggesting that algorithm selection has a more pronounced impact when the classification task is inherently harder.

4 COMPUTATIONAL EFFICIENCY ANALYSIS

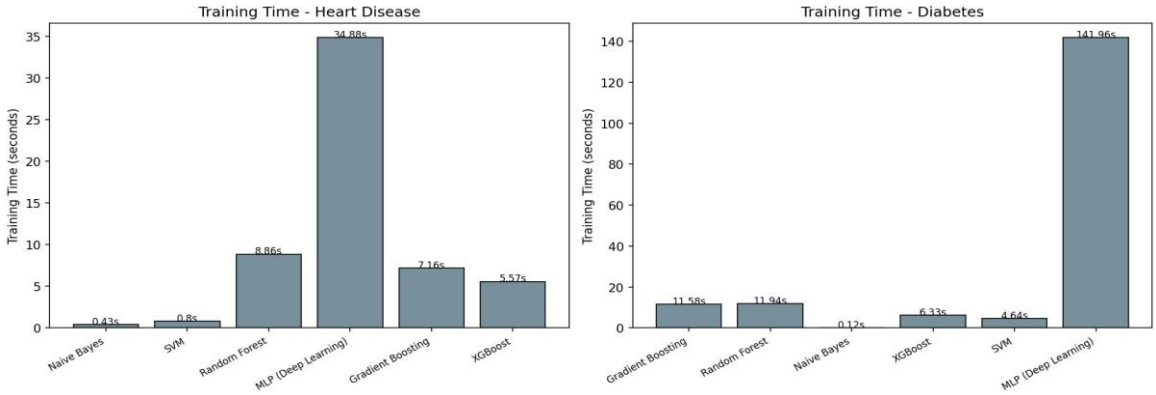
Figure 4 summarizes the training times for all classifiers. Naive Bayes is the fastest model by a substantial margin (0.43s on Heart Disease, 0.12s on Diabetes), owing to its closed-form parameter estimation. SVM trains quickly on

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

small datasets but would be expected to scale poorly to very large datasets due to its O(n²) to O(n³) complexity. Random Forest and Gradient Boosting show moderate training times (7–12 seconds), representing a practical balance between performance and efficiency.

Figure 4. Training Time Comparison — Heart Disease (left) and Diabetes (right)



MLP exhibits the highest training times (34.88s on Heart Disease, 141.96s on Diabetes), which scale poorly with data volume. In a Big Data context where model retraining is frequent, this computational overhead must be carefully weighed against marginal performance gains — particularly given that MLP did not outperform simpler models in our experiments.

5 DISCUSSION AND PRACTICAL RECOMMENDATIONS

The experimental results lead to several practical recommendations for the design of health information systems:

For resource-constrained environments: Naive Bayes offers an excellent accuracy-efficiency trade-off, particularly for cardiovascular disease prediction. Its near-instantaneous training time makes it highly suitable for real-time clinical decision support systems and edge computing scenarios in healthcare IoT.

For maximum predictive performance: Gradient Boosting and Random Forest consistently deliver the best overall performance across both datasets and metrics. These ensemble methods are recommended as default classifiers for health prediction tasks, especially when recall on the positive class is clinically critical.

For scalability in Big Data environments: XGBoost provides a strong balance of performance and computational efficiency, with built-in distributed training support (via Dask or Spark integration), making it particularly well-suited for large-scale health data processing pipelines.

Regarding deep learning: MLP underperformed on both small datasets while incurring the highest computational cost. Deep learning approaches are not recommended for small-to-medium sized medical datasets without data augmentation or transfer learning strategies.

These findings are consistent with the broader literature showing that algorithm performance is highly context-dependent, and that no single model universally dominates across all health informatics tasks [7].

CONCLUSION

This paper presented a systematic comparative study of six machine learning algorithms for disease prediction in public health, evaluated within a Big Data information system framework. Experiments on two well-known benchmark datasets — the UCI Heart Disease dataset and the Pima Indians Diabetes dataset — using 10-fold stratified cross-validation demonstrated that:

(1) Naive Bayes achieved the best accuracy (83.78%) and AUC-ROC (90.03%) on the Heart Disease dataset, combining excellent predictive performance with near-instantaneous training time.

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

(2) Gradient Boosting led on the Diabetes dataset with an accuracy of 77.72% and an AUC-ROC of 83.21%, confirming the superiority of ensemble methods for complex, imbalanced health classification tasks.

(3) MLP, despite its architectural sophistication, did not outperform simpler models on either dataset, while incurring substantially higher computational costs — a critical limitation for Big Data health applications.

(4) The proposed Big Data pipeline provides a scalable and reproducible framework for integrating machine learning into health information systems, with practical guidance for algorithm selection based on performance-efficiency trade-offs.

Future work will focus on three directions: (i) extending the framework to larger real-world clinical datasets processed through Apache Spark, (ii) investigating federated learning approaches to address patient privacy constraints in distributed health data environments, and (iii) incorporating explainability mechanisms (e.g., SHAP values) to support clinical interpretability of model predictions.

REFRENCES

[1] Dash, S., Shakyawar, S. K., Sharma, M., and Kaushik, S. (2019). Big data in healthcare: Management, analysis and future prospects. Journal of Big Data, 6(1), 54.
[2] Belle, A., Thiagarajan, R., Soroushmehr, S. M. R., Navidi, F., Beard, D. A., and Najarian, K. (2015). Big data analytics in healthcare. BioMed Research International, 2015, 370194.
[3] Obermeyer, Z., and Emanuel, E. J. (2016). Predicting the future — big data, machine learning, and clinical medicine. New England Journal of Medicine, 375(13), 1216–1219.
[4] Palaniappan, S., and Awang, R. (2008). Intelligent heart disease prediction system using data mining techniques. In IEEE/ACS International Conference on Computer Systems and Applications (pp. 108–115).
[5] Raghupathi, W., and Raghupathi, V. (2014). Big data analytics in healthcare: Promise and potential. Health Information Science and Systems, 2(1), 3.
[6] Saravanan, R., and Pabitha, P. (2019). A state of art techniques on machine learning algorithms: A perspective of supervised learning approaches in data classification. In Proceedings of the Second International Conference on Intelligent Computing and Control Systems.
[7] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., and Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Computational and Structural Biotechnology Journal, 15, 104–116.
[8] Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., and Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. Frontiers in Genetics, 9, 515.
[9] Mohan, S., Thirumalai, C., and Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. IEEE Access, 7, 81542–81554.
[10] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.
[11] Chen, T., and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794).