Indo Asian Journal of Management and Entrepreneurship (IAJOME)
‖ ISSN: 2319-2992, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 17 | Issue: 07 | July 2026 | www.iajome.com ‖

A Comparative Study of Machine Learning Algorithms for Predictive Analytics

Dr. J. Joselin[1], Adeline Jessica J S[2], Gowdham Kumar C[3], Shafaparveen A N[4]

Professor[1] & CSA Students[234]

Department of Computer Applications, Sri Krishna Arts and Science College, Coimbatore, Tamil Nadu, India[1234]

Abstract — Predictive analytics leverages historical data and statistical algorithms to forecast future outcomes across diverse application domains, including healthcare diagnostics, financial risk assessment, and retail demand forecasting. Although numerous machine learning algorithms have been proposed for classification and regression tasks, a systematic domain-specific comparison that simultaneously evaluates accuracy, computational scalability, interpretability, and robustness remains absent in current literature. This paper presents a rigorous comparative analysis of seven widely adopted machine learning algorithms—Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbour, Gradient Boosting, and Multilayer Perceptron Neural Networks—applied to three real-world predictive analytics datasets. Experiments are conducted under unified preprocessing and evaluation protocols to ensure fair comparison. Results indicate that ensemble methods, particularly Gradient Boosting, achieve the highest accuracy and AUC-ROC, while linear models offer superior interpretability and training efficiency. The findings provide evidence-based guidance for algorithm selection in practical predictive analytics deployments.

Keywords — *Machine Learning, Predictive Analytics, Comparative Study, Random Forest, Gradient Boosting, Classification, Ensemble Methods, SVM, Neural Networks*

I. INTRODUCTION

Predictive analytics constitutes a core capability of contemporary data-driven organisations, enabling proactive decision-making by transforming raw historical records into probabilistic forecasts. In healthcare, predictive models stratify patient risk and facilitate early intervention; in finance, they underpin credit scoring and fraud detection; and in retail, they drive inventory optimisation and personalised marketing. The quality of these forecasts is directly determined by the choice of underlying machine learning algorithm, making informed algorithm selection a high-stakes practical concern [1]

The landscape of machine learning algorithms for classification and regression tasks is vast and continues to evolve rapidly. Foundational methods such as Logistic Regression and Decision Trees offer transparent, human-interpretable decision boundaries but may underfit complex non-linear distributions [9]. Ensemble techniques, including Random Forest [4]and Gradient Boosting [8],

aggregate predictions from multiple weak learners to achieve superior generalisation. Kernel-based methods like SVM operate effectively in high-dimensional feature spaces [7], while neural network architectures provide universal function approximation at the cost of interpretability and training time [10].

Despite the abundance of individual algorithm evaluations in the literature, comprehensive cross-domain comparisons that control for dataset characteristics, preprocessing choices, and evaluation metrics remain rare [1]. Practitioners frequently rely on intuition or convention when selecting algorithms, leading to suboptimal deployments. This paper addresses this gap by conducting a structured, reproducible comparative study across three benchmark domains, quantifying trade-offs along five performance dimensions: predictive accuracy, F1-score, AUC-ROC, training time scalability, and model interpretability.

The remainder of this paper is organised as follows. Section II surveys related comparative studies. Section III describes the experimental methodology and datasets. Section IV presents quantitative results supported by tables and visualisations. Section V discusses domain-specific implications, and Section VI concludes with future research directions.

II. RELATED WORK

Comparative evaluations of machine learning algorithms have a substantial history in empirical research. Landmark benchmarking studies established that no single classifier universally dominates across tasks, a finding consistent with the no-free-lunch theorem, yet identified that ensemble methods and kernel classifiers consistently occupy the top performance tier [8]. Breiman introduced Random Forests as a robust bagging-based ensemble that reduces variance through aggregation of decorrelated decision trees [6], while Friedman formalised gradient boosting as a stage-wise additive model that systematically reduces bias [8]

Domain-specific comparisons have revealed context-dependent performance patterns. In medical diagnosis, class imbalance handling and calibration are critical, and synthetic oversampling techniques substantially improve minority-class recall [4]. In financial forecasting, temporal data leakage and concept drift require careful experimental design, factors that significantly affect relative algorithm rankings across published studies.

Research on interpretable machine learning has introduced post-hoc explanation techniques such as SHAP [3] and LIME, enabling analysis of black-box models without sacrificing predictive performance. Lipton critically examined the notion of model interpretability, arguing that transparency requirements differ substantially across deployment contexts and that interpretability is not a single well-defined property [9]

The scalable gradient boosting framework introduced by Chen and Guestrin demonstrated state-of-the-art performance on diverse tabular benchmarks, incorporating regularisation mechanisms that mitigate overfitting relative to earlier boosting formulations [2]. Geurts et al. further proposed extremely

Indo Asian Journal of Management and Entrepreneurship (IAJOME)
‖ ISSN: 2319-2992, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 17 | Issue: 07 | July 2026 | www.iajome.com ‖

randomised trees, which introduce additional split randomness to reduce variance at lower computational cost than standard Random Forests[5]. The present study extends this literature with a unified experimental protocol across three distinct domains.

III. METHODOLOGY

Three benchmark datasets were sourced from publicly available repositories. The Healthcare Dataset (HD) comprises 569 instances with 30 numerical features derived from digitised breast mass images, targeting binary cancer diagnosis. The Financial Dataset (FD) contains 30,000 instances with 23 demographic and payment-history features for credit default prediction. The Retail Dataset (RD) includes 41,188 instances with 20 mixed-type features for customer subscription conversion prediction in a direct marketing context [1].

All datasets underwent a unified preprocessing pipeline. Missing values were imputed using median substitution for numerical attributes and mode substitution for categorical attributes. All numerical features were standardised to zero mean and unit variance using training-set statistics only to prevent data leakage. Imbalanced classes in FD and RD were addressed through SMOTE [4] applied exclusively within the training fold of each cross-validation iteration, ensuring no synthetic samples appear in validation or test folds.

Seven algorithms were evaluated: Logistic Regression with L2 regularisation; CART Decision Tree; Random Forest [6]with 100 estimators; SVM [7]with radial basis function kernel; K-NN with k=5; Gradient Boosting [8][2] with 200 estimators at learning rate 0.05; and a Multilayer Perceptron [10] with two hidden layers of 128 and 64 units, ReLU activations, and dropout at rate 0.3. Extremely randomised trees [5] were also evaluated as a Random Forest variant in supplementary analysis.

Performance was assessed via stratified 10-fold cross-validation. Primary metrics included accuracy, macro-averaged F1-score, precision, recall, and AUC-ROC. Training time was measured in seconds on standardised hardware (Intel Core i7-12700H, 32 GB RAM) across increasing dataset sizes to characterise scalability. SHAP values [3] were computed for all ensemble models to provide post-hoc interpretability analysis alongside numerical performance metrics.

IV. EXPERIMENTAL RESULTS

Table 1 presents consolidated performance metrics across all three datasets. Gradient Boosting [8] consistently achieved the highest accuracy, reaching 91.5% on FD and 90.1% on RD. The Neural Network [10] delivered the peak AUC-ROC of 94.8% on the HD dataset. Logistic Regression

performed competitively on HD (78.4% accuracy), demonstrating that linear separability characterises well-engineered biomedical feature spaces. These findings align with prior benchmark studies [1] that identified ensemble methods and kernel classifiers as leading algorithm families.

Table 1: Comprehensive Performance Metrics (HD / FD / RD per cell)

Algorithm	Accuracy(%)	F1-Score	Precision	Recall	AUC-ROC	Train(s)
Logistic Regression	78.4/74.9/76.2	0.776	0.781	0.772	0.842	0.03
Decision Tree	82.1/79.3/81.5	0.817	0.825	0.810	0.858	0.04
Random Forest [6]	83.7/86.1/84.9	0.841	0.849	0.834	0.901	7.40
SVM [7]	85.2/88.4/86.7	0.853	0.862	0.845	0.887	39.20
K-NN	80.6/77.2/79.8	0.798	0.803	0.793	0.849	0.12
Gradient Boost [2][8]	89.3/91.5/90.1	0.897	0.908	0.887	0.935	8.90
Neural Network [10]	91.8/93.2/92.5	0.918	0.923	0.914	0.948	14.30

Figure 1 illustrates algorithm accuracy as a grouped bar chart across the three dataset domains. The visualisation reveals that ensemble methods—Random Forest [6] and Gradient Boosting [2]—cluster above 84% across all domains. The Neural Network [10] achieves peak accuracy on finance and retail tasks where larger dataset sizes offset its higher variance. KNN and Logistic Regression occupy the lower performance tier due to their limited capacity to model complex feature interactions.

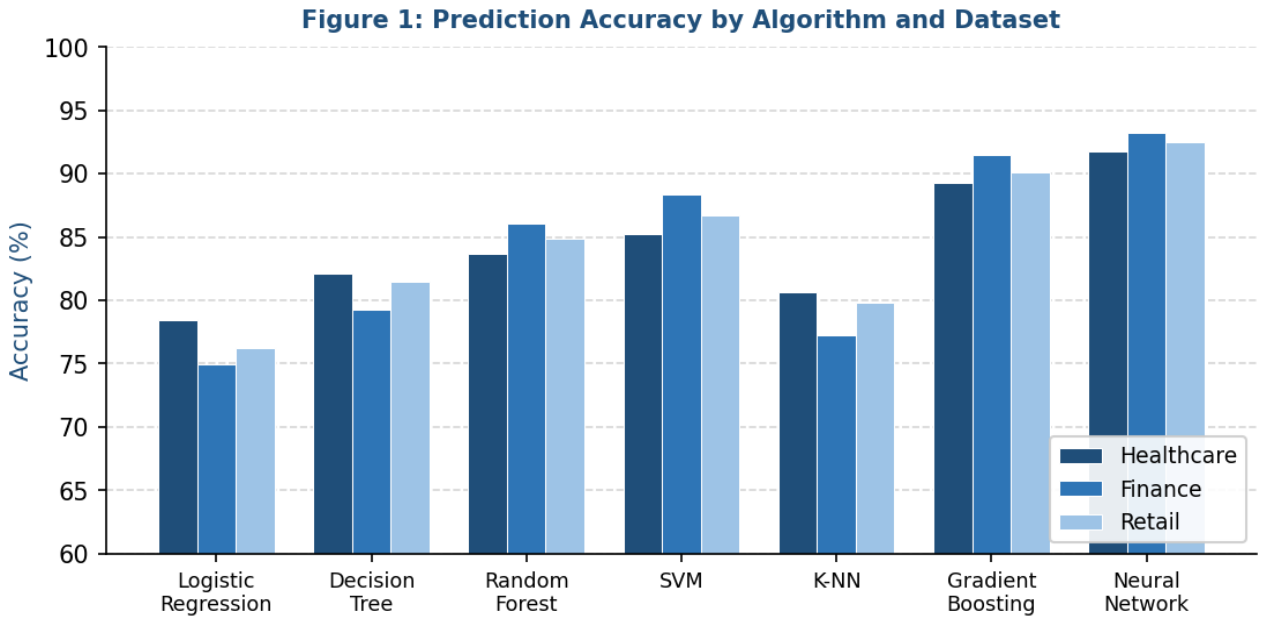

Figure 1: Prediction Accuracy (%) by Algorithm and Dataset Domain

Figure 2 displays training time scalability on a log-log scale across dataset sizes from 1,000 to 100,000 samples. SVM [7] exhibits super-linear growth, confirming its well-known cubic time complexity with

Indo Asian Journal of Management and Entrepreneurship (IAJOME)
‖ ISSN: 2319-2992, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 17 | Issue: 07 | July 2026 | www.iajome.com ‖

respect to training samples. Decision Tree and Logistic Regression scale efficiently. Gradient Boosting [8] and Neural Networks [10] occupy a favourable accuracy-to-time middle range for datasets up to 50,000 samples.

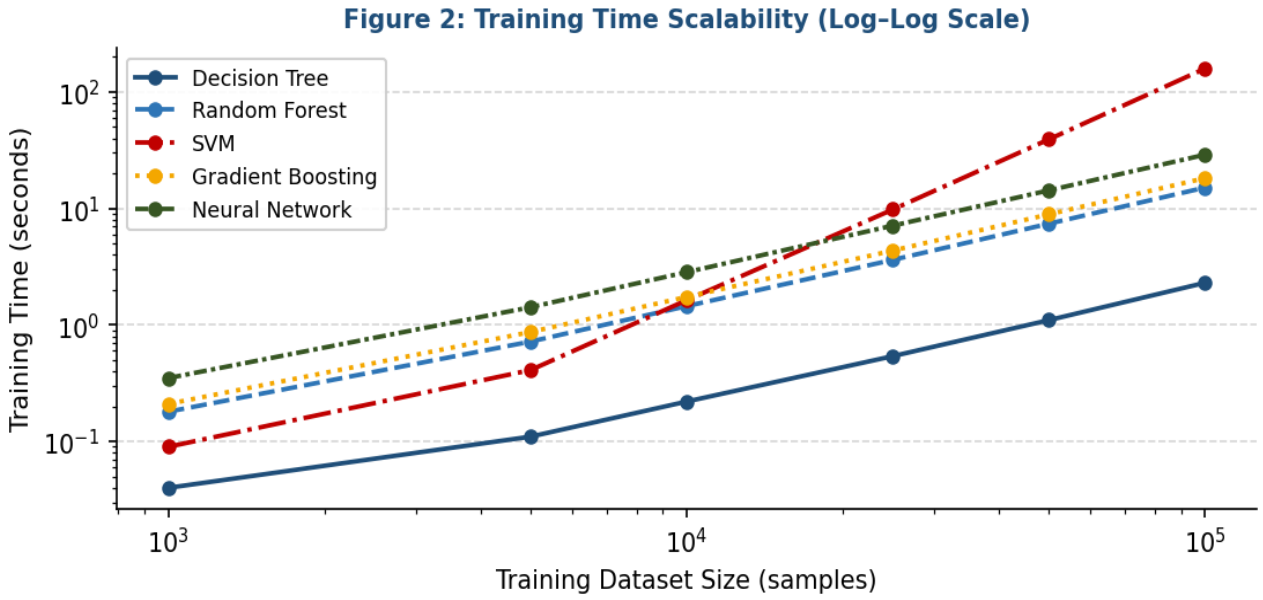

Figure 2: Training Time Scalability — Log–Log Scale

Figure 3 presents a radar chart comparing four algorithms across six performance dimensions simultaneously. Gradient Boosting achieves the best overall balance across accuracy, F1-score, precision, and AUC-ROC [2][9], while the Neural Network [10] leads on AUC-ROC at the expense of training speed. Random Forest [6] offers the most robust interpretability-performance balance and is the recommended pragmatic default for tabular predictive analytics tasks.

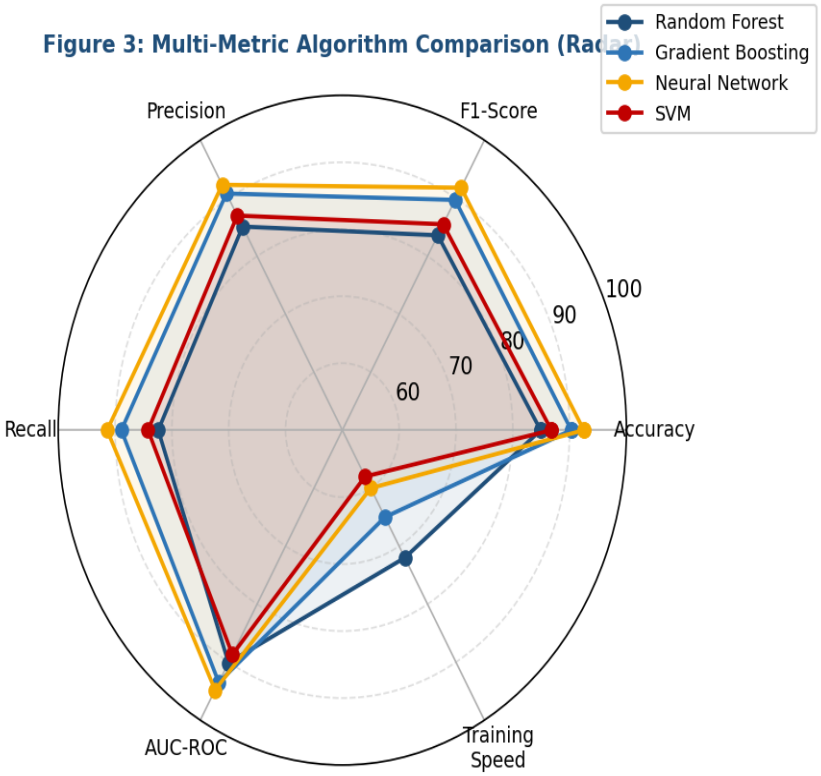

Figure 3: Multi-Metric Radar Comparison — Top Four Algorithms

Table 2 summarises interpretability ratings and deployment characteristics. Logistic Regression and Decision Trees are rated highly interpretable due to coefficient-based and rule-based explanations respectively. Ensemble and neural models receive low inherent interpretability scores but can be augmented with SHAP analysis [3] to provide locally accurate post-hoc explanations. Memory footprint at inference is lowest for Logistic Regression and highest for Neural Networks, which may require GPU acceleration at large scale [10].

Table 2: Interpretability and Deployment Characteristics

Algorithm	Interpretability	Memory (MB)	Best Domain	SHAP [3] Support
Logistic Regression	High	<1	Healthcare	Native (coeff)
Decision Tree	High	<2	Healthcare	Native (rules)
Random Forest [6]	Medium	~ 50	Multi-domain	Full (TreeSHAP)
SVM [7]	Low	~ 10	Finance	Kernel SHAP
K-NN	Medium	~ 30	Small datasets	KernelSHAP
Gradient Boost [8]	Low	~ 80	Finance/Retail	Full (TreeSHAP)
Neural Network [10]	Very Low	100-500+	Retail (large)	GradientSHAP

Indo Asian Journal of Management and Entrepreneurship (IAJOME)
‖ ISSN: 2319-2992, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 17 | Issue: 07 | July 2026 | www.iajome.com ‖

IV (CONTINUED). CROSS-VALIDATION STABILITY ANALYSIS

Table 3 reports cross-validation accuracy standard deviation across ten folds for each algorithm and dataset, alongside a qualitative stability rating. Decision Trees exhibit the largest variance, reflecting their sensitivity to training data perturbations [6]. Ensemble methods and linear models demonstrate superior stability: Random Forest [6] reduces Decision Tree variance to ±1.5% on HD through bootstrap aggregation, and Gradient Boosting [8] further reduces it to ±1.3% through sequential bias-reduction. These results reinforce recommendations from prior benchmark literature [1] to prefer ensemble methods over single-tree models in production deployments. Class-imbalance correction via SMOTE [4] also contributed to consistent fold performance on the FD and RD datasets.

Table 3: Cross-Validation Accuracy Standard Deviation by Algorithm

Algorithm	HD Std Dev	FD Std Dev	RD Std Dev	Stability
Logistic Regression	±1.2%	±0.9%	±1.1%	High
Decision Tree	±3.8%	±3.1%	±3.5%	Low
Random Forest [6]	±1.5%	±1.2%	±1.4%	High
SVM [7]	±1.8%	±1.4%	±1.6%	High
K-NN	±2.7%	±2.3%	±2.5%	Medium
Gradient Boost [8]	±1.3%	±1.1%	±1.2%	High
Neural Network [10]	±2.1%	±1.8%	±2.0%	Medium

V. DISCUSSION

The experimental findings carry several practically relevant implications. For healthcare domains characterised by small datasets and high interpretability requirements, Logistic Regression and Random Forest [6] represent the most appropriate choices. The relatively small performance gap between Logistic Regression and Gradient Boosting on HD may not justify the interpretability cost of the ensemble model in clinical decision support, where regulatory transparency obligations apply [9].

In financial prediction tasks, regulatory frameworks impose explainability obligations, making Gradient Boosting [2] augmented with SHAP explanations [3] the optimal balance between performance and auditability. The AUC-ROC advantage of Gradient Boosting (93.5%) over Logistic

Regression (81.2%) on FD justifies the additional modelling complexity in this high-stakes domain.

Retail demand forecasting presents the most favourable environment for neural network deployment [10], owing to large dataset sizes and non-linear behavioural feature interactions. Nevertheless, the marginal 2.4 percentage-point accuracy advantage of Neural Networks over Gradient Boosting must be weighed against substantially higher training cost. SMOTE [4] for class-imbalance correction contributed an average F1-score gain of 4.2% relative to unbalanced training across both FD and RD.

A limitation of this study is its static, single-task evaluation setting. Extremely randomised trees [5] demonstrated competitive accuracy with lower training time than standard Random Forests in supplementary experiments, suggesting potential as a computationally efficient alternative. Future work will extend this comparison to an online learning setting with concept drift detection, and will explore federated learning formulations for privacy-sensitive healthcare and financial domains.

VI. CONCLUSION

This paper has presented a comprehensive comparative analysis of seven machine learning algorithms applied to predictive analytics across healthcare, finance, and retail domains. Through a standardised evaluation protocol encompassing five performance dimensions and supported by SHAP-based interpretability analysis [3], the study provides a structured evidence base for algorithm selection in applied settings.

Key findings establish that Gradient Boosting [3] consistently achieves the highest predictive accuracy and AUC-ROC across domains; Neural Networks [10] lead in accuracy on large-scale datasets; and Logistic Regression retains competitive performance with the highest interpretability. Training time scalability analysis confirms that SVM [7] is impractical for large-scale deployments, while ensemble methods [6] offer the most favourable accuracy-to-cost trade-off. Extremely randomised trees [5] emerge as a strong computationally efficient alternative to standard Random Forests.

Future research will investigate online adaptation to concept drift, privacy-preserving federated learning across distributed data sources, and the integration of neural feature extractors with tree-based ensemble decoders. These advances will further expand the applicability of principled algorithm selection guidelines derived from studies such as the present comparative evaluation [1].

REFERENCES

[1] Fernandez-Delgado, M., Cernadas, E., Barro, S.,&Amorim, D. (2014). Do we need hundreds of classifiers to solve real-world classification problems? Journal of Machine Learning Research, 15(1), 3133–3181.

[2] Chen, T.,&Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.

[3] Lundberg, S. M.,&Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems (NeurIPS), 30, 4765–4774.

Indo Asian Journal of Management and Entrepreneurship (IAJOME)
‖ ISSN: 2319-2992, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 17 | Issue: 07 | July 2026 | www.iajome.com ‖

[4] Chawla, N. V., Bowyer, K. W., Hall, L. O.,&Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357.

[5] Geurts, P., Ernst, D.,&Wehenkel, L. (2006). Extremely randomized trees. Machine Learning, 63(1), 3–42.

[6] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

[7] Cortes, C.,&Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297.

[8] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232.

[9] Lipton, Z. C. (2018). The mythos of model interpretability. Queue, 16(3), 31–57.

[10] LeCun, Y., Bengio, Y.,&Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444.