VIVA-Tech International Journal for Research and Innovation
ISSN(Online): 2581-7280

Volume 1, Issue 9 (2026)
PP 01-06

VIVA-TECH IJRI
ANNUAL RESEARCH JOURNAL

VIVA-TECH INTERNATIONAL JOURNAL FOR RESEARCH AND INNOVATION
ANNUAL RESEARCH JOURNAL
ISSN(ONLINE): 2581-7280

A Comparative Study of Machine Learning Algorithms for Predictive Data Analysis

Sonia Dubey[1], Satyam Chavan[2], Vignesh Asolkar[3]
[1](Department of Computer Applications, VIVA Institute of Technology, India)
[2](Department of Computer Applications, VIVA Institute of Technology, India)
[3](Department of Computer Applications, VIVA Institute of Technology, India)

Abstract

Machine learning has become essential for predictive data analysis across diverse domains including healthcare, finance, and manufacturing. Despite the availability of numerous algorithms, practitioners face challenges in selecting appropriate methods for specific analytical tasks. This paper presents a qualitative analysis of machine learning algorithm selection criteria and proposes a structured decision framework based on data characteristics, computational constraints, and interpretability requirements. Through comprehensive literature review, we identify key factors influencing algorithm performance and provide practical guidance for algorithm selection without requiring extensive experimental comparison

Keywords

Machine learning, algorithm selection, predictive analytics, decision framework, data analysis

1. Introduction

Predictive data analysis has emerged as a critical capability for organizations seeking to leverage data-driven insights for decision-making. Machine learning algorithms enable systems to identify patterns, make predictions, and generate actionable intelligence from the historical data. The landscape of available algorithms spans simple linear models to complex deep neural networks, each with distinct characteristics, advantages, and limitations.

Practitioners frequently encounter the challenge of algorithm selection when initiating predictive analytics projects. The abundance of available methods, combined with varying data characteristics and project constraints, creates uncertainty regarding optimal choices. Current approaches to algorithm selection often rely on trial-and-error experimentation, benchmark comparisons, and heuristic rules developed through experience. These methods consume significant computational resources and time.

This study addresses the gap between theoretical algorithm knowledge and practical selection decisions. Through a qualitative analysis of the existing literature and algorithmic properties, we developed a structured framework that guides practitioners toward appropriate algorithm choices based on observable data characteristics and the project requirements.

Volume 1, Issue 9 (2026)
PP 01-06

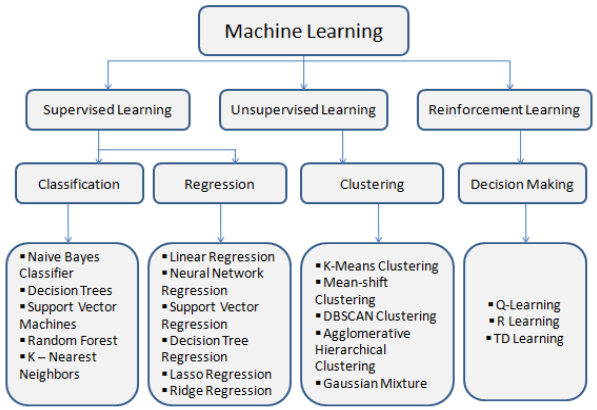

Figure 1: Overview of major machine learning algorithm categories for predictive analytics

2. Literature Review

Author & Year	Key Findings	Conclusion
Fernández-Delgado et al. (2014)	Evaluated 179 classifiers across 121 datasets; no single algorithm dominates all scenarios	Algorithm performance is highly context-dependent; selection must consider data characteristics
Caruana & Niculescu-Mizil (2006)	Compared supervised learning algorithms; ensemble methods show consistent performance	Ensemble approaches provide robust performance across diverse tasks
Wolpert & Macready (1997)	Established "no free lunch" theorem for optimization	Universal superiority of any algorithm is mathematically impossible; context matters
Domingos (2012)	Analyzed fundamental tradeoffs in machine learning	Algorithm selection involves balancing multiple competing objectives
Hand (2006)	Examined classifier performance measurement challenges	Performance metrics must align with application objectives
Kotsiantis et al. (2007)	Reviewed supervised learning algorithms and applications	Different domains favor different algorithmic approaches

The literature reveals several consistent themes regarding the selection of algorithms. First, no universal algorithm is the best across all problem domains and data characteristics. Second, the algorithm performance depends critically on factors such as dataset size, feature dimensionality, noise levels, and complexity of the underlying patterns. Third, practical constraints, such as computational resources, interpretability requirements, and deployment considerations, significantly influence the suitability of algorithms.

Research on meta-learning and algorithm selection has identified data characteristics that correlate with the performance of algorithms. These include statistical properties (mean, variance, skewness), information-theoretic measures (entropy, mutual information), and structural features (class balance and feature correlation). However, translating these insights into actionable selection frameworks remains challenging for practitioners.

Studies have emphasized the importance of matching algorithm capabilities to problem requirements. Linear models excel when relationships are approximately linear, the data are limited, or interpretability is paramount. Tree-based methods handle nonlinear patterns and mixed data types and provide feature importance naturally. Support vector machines perform well in high-dimensional spaces with clear margins. Neural networks require substantial amounts of data but can learn complex hierarchical representations.

The trade-off between model complexity and interpretability is a critical consideration, particularly in regulated domains such as healthcare and finance. Although complex models may achieve higher predictive accuracy, simpler interpretable models facilitate validation, debugging, and regulatory compliance.

3. Problem Definition

This study adopts a qualitative methodology based on a comprehensive literature review and systematic analysis of algorithm properties. This approach synthesizes insights from theoretical research, empirical studies, and documented practical experiences.

The methodology comprises several components, as follows. First, an extensive review of the academic literature on machine learning algorithms, including theoretical foundations, empirical comparisons, and application studies, was conducted. Second, a systematic analysis of algorithm characteristics across multiple dimensions, including computational complexity, data requirements, interpretability and scalability. Third, the real-world constraints affecting algorithm deployment must be examined.

No original data were collected, and no algorithms were implemented. This research relies entirely on the synthesis of existing knowledge, making it suitable for organizations seeking guidance without conducting original experiments

4. Objectives

- To systematically analyze machine learning algorithm characteristics relevant to selection decisions
- To identify observable data properties that influence algorithm suitability
- To examine practical constraints affecting algorithm selection in real-world contexts
- To develop a structured decision framework for algorithm selection based on qualitative criteria
- To provide actionable guidance for practitioners facing algorithm selection challenges

5. Methodology

This study adopts a qualitative methodology based on a comprehensive literature review and systematic analysis of algorithm properties. This approach synthesizes insights from theoretical research, empirical studies, and documented practical experiences.

The methodology comprises several components, as follows. First, an extensive review of the academic literature on machine learning algorithms, including theoretical foundations, empirical comparisons, and application studies, was conducted. Second, a systematic analysis of algorithm characteristics across multiple dimensions, including computational complexity, data requirements, interpretability and scalability. Third, the real-world constraints affecting algorithm deployment must be examined.

No original data were collected, and no algorithms were implemented. This research relies entirely on the synthesis of existing knowledge, making it suitable for organizations seeking guidance without conducting original experiments.

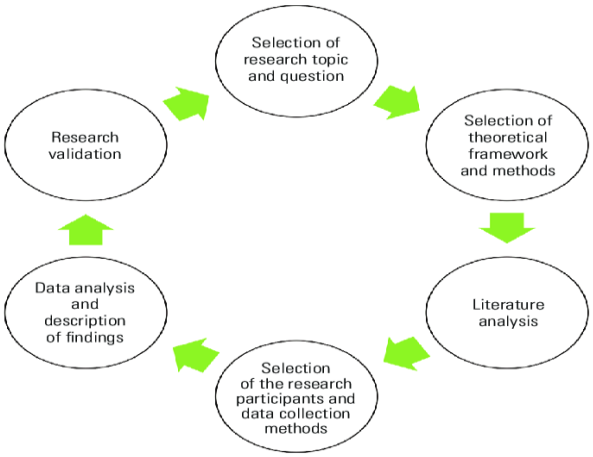

Figure 2: Research methodology process for developing the selection framework

6. Results and Discussion

6.1 Algorithm Characteristics Analysis

Qualitative analysis revealed distinct characteristic profiles for the major algorithm categories.

Linear Models: Computational efficiency, interpretability through coefficients, limited capacity for complex patterns, minimal tuning required, and low sample size requirements.

Decision Trees: Natural nonlinearity handling, interpretable rules, overfitting tendency, feature importance calculation, and efficient computation.

Random Forests: Reduced overfitting through ensembles, robust performance, moderate computational needs, minimal hyperparameter sensitivity, and strong default performance.

Support Vector Machines: Effectiveness in high dimensions, strong generalization, kernel flexibility, computational scaling challenges, and hyperparameter sensitivity.

Neural Networks: High capacity for complexity, automatic feature learning, large data requirements, substantial computational demands, and limited interpretability.

6.2 Proposed Selection Framework

Based on this analysis, we propose a structured decision framework.

Step 1: Assess Interpretability Requirements

- High interpretability needed → Linear models or decision trees
- Moderate interpretability → Random forests
- Low interpretability constraint → All algorithms considered

Step 2: Evaluate Dataset Size

- Small (<1,000 samples) → Simple models (linear, regularized)
- Medium (1,000-10,000) → Random forests, SVMs
- Large (>10,000) → Neural networks feasible

Step 3: Consider Computational Resources

- Limited → Linear models, decision trees
- Moderate → Random forests, SVMs
- Substantial → Neural networks, extensive tuning

Step 4: Apply Validation Strategy

- Cross-validation for performance estimation
- Hold-out test set for final evaluation
- Compare against simple baselin

Framework Application Example

Scenario: Healthcare predictive model for patient readmission risk

- Interpretability: High (regulatory compliance required) → Step 2a
- Relationships: Potentially non-linear → Decision trees recommended
- Dataset size: 5,000 patients → Adequate for decision trees
- Selection: Decision tree with limited depth, validated through cross-validation

Scenario: E-commerce product recommendation

- Interpretability: Low (internal system) → Step 2c
- Dataset size: 500,000 interactions → Large dataset
- Computational resources: Substantial → Neural networks feasible
- Selection: Neural network with collaborative filtering architecture

Scenario: Manufacturing quality prediction

- Interpretability: Moderate (explanation helpful but not mandatory) → Step 2b
- Dataset size: 10,000 measurements → Medium dataset
- Time constraint: Moderate → Random forests
- Selection: Random forest with default parameters, validate and tune if needed

Figure 3: Example applications of the framework across different industries

6.3 Comparative Analysis Summary

Algorithm	Best Use Cases	Key Advantage	Main Limitation
Linear Models	Linear patterns, interpretability needs	Speed & clarity	Limited capacity
Decision Trees	Mixed data, transparency required	Interpretability	Overfitting
Random Forests	General purpose, medium data	Versatility	Moderate complexity
SVM	High dimensions, clear margins	Generalization	Scaling issues

Algorithm	Best Use Cases	Key Advantage	Main Limitation
Neural Networks	Large data, complex patterns	Highest capacity	Data hungry

7. Conclusion and Future Work

This qualitative analysis provides a structured framework for machine-learning algorithm selection based on assessable data characteristics and practical constraints. The key findings confirmed that no universal best algorithm exists; the selection must account for specific data properties, including dataset size, feature dimensionality, and interpretability requirements.

The proposed framework enables practitioners to make informed selections without extensive experimentation, thereby saving computational resources and accelerating the project timelines. The advantages include practical applicability and systematic organization of existing knowledge into actionable guidance. The limitations include the qualitative nature, which precludes precise performance predictions, and the focus on major algorithm categories rather than specific variants.

Future work should pursue empirical validation through case studies, integration with automated machine learning tools, and extension to additional algorithm families, including gradient boosting and domain-specific applications such as time-series forecasting and natural language processing.

8. Acknowledgements

The authors express sincere gratitude to the Department of Master of Computer Applications and Viva Institute of Technology for providing academic support and resources necessary for this research.

9. References

[1] C. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" Journal of Machine Learning Research, vol. 15, pp. 3133-3181, 2014. https://jmlr.org/papers/v15/delgado14a.html
[2] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," Proceedings of the 23rd International Conference on Machine Learning, pp. 161-168, 2006. https://dl.acm.org/doi/10.1145/1143844.1143865
[3] D. Wolpert and W. Macready, "No free lunch theorems for optimization," IEEE Transactions on Evolutionary Computation, vol. 1, no. 1, pp. 67-82, 1997. https://ieeexplore.ieee.org/document/585893
[4] P. Domingos, "A few useful things to know about machine learning," Communications of the ACM, vol. 55, no. 10, pp. 78-87, 2012. https://dl.acm.org/doi/10.1145/2347736.2347755
[5] D. Hand, "Classifier technology and the illusion of progress," Statistical Science, vol. 21, no. 1, pp. 1-14, 2006. https://projecteuclid.org/euclid.ss/1149600839
[6] S. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques," Emerging Artificial Intelligence Applications in Computer Engineering, vol. 160, pp. 3-24, 2007. https://www.researchgate.net/publication/228084860
[7] R. Caruana, A. Niculescu-Mizil, G. Crew, and A. Ksikes, "Ensemble selection from libraries of models," Proceedings of the 21st International Conference on Machine Learning, 2004. https://dl.acm.org/doi/10.1145/1015330.1015432
[8] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," International Joint Conference on Artificial Intelligence, vol. 14, pp. 1137-1145, 1995. https://www.ijcai.org/Proceedings/95-2/Papers/016.pdf
[9] P. Brazdil, C. Giraud-Carrier, C. Soares, and R. Vilalta, "Metalearning: Applications to data mining," Springer Science & Business Media, 2009. https://link.springer.com/book/10.1007/978-3-540-73263-1
[10] C. Lemke, M. Budka, and B. Gabrys, "Metalearning: A survey of trends and technologies," Artificial Intelligence Review, vol. 44, no. 1, pp. 117-130, 2015. https://link.springer.com/article/10.1007/s10462-013-9406-y
[11] S. Ali and K. Smith-Miles, "A meta-learning approach to automatic kernel selection for support vector machines," Neurocomputing, vol. 70, no. 1-3, pp. 173-186, 2006. https://www.sciencedirect.com/science/article/abs/pii/S0925231206002773