IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, May 2025
Impact Factor: 7.67

A Systematic Review of Multi-Class Classification Techniques Across Diverse Domains

[1]Tejas Moon, [2]Prof. Dr. Bireshwar Ganguly, [3]Prof. Dr. Devashri Raich
[1]M.Tech (CSE), Final Year, Department of Computer Science and Engineering
[2,3]Assistant Professor, Department of Computer Science and Engineering
Rajiv Gandhi College of Engineering, Research and Technology, Chandrapur

Abstract: *This review examines the comparative performance of supervised machine learning algorithms for multi-class classification across diverse domains such as healthcare, social media analysis, and benchmark datasets. A structured literature search was conducted to identify studies that empirically evaluate and compare models including Random Forest, Gradient Boosting, Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Naïve Bayes, Multinomial Logistic Regression, and Decision Tree. The findings indicate that no single algorithm consistently outperforms others; instead, performance depends on dataset characteristics, class distribution, and preprocessing techniques such as feature selection and normalization. Evaluation practices commonly rely on metrics like accuracy, precision, recall, and F1-score, though these may be insufficient for imbalanced datasets. The review highlights current methodological limitations and proposes future directions, including hybrid model development, transfer learning, and explainable AI. These insights aim to guide researchers and practitioners in selecting and optimizing classification models for multi-class problems.*

Keywords: Supervised machine learning, algorithm comparison, multi-class classification, performance evaluation

I. INTRODUCTION

Multi-class classification is a core task in supervised machine learning, aiming to assign each data instance to one of three or more specified categories. This form of classification is essential in numerous fields, such as disease diagnosis, sentiment analysis, image recognition, and other areas where decisions rely on complex, multi-label outputs. As datasets become more intricate, with diverse feature distributions, class imbalances, and domain-specific issues, the effectiveness of classification algorithms becomes increasingly dependent on the nature of the data and the preprocessing methods employed.

A variety of supervised learning algorithms have been utilized for multi-class classification, each bringing unique advantages. Methods like Random Forest, Gradient Boosting, Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Naïve Bayes, Multinomial Logistic Regression, and Decision Tree have been widely researched in both theoretical and practical settings. The relative performance of these algorithms is often influenced by factors such as class balance, data dimensionality, and the use of preprocessing steps like normalization and feature selection.

This review consolidates recent research that empirically assesses and contrasts the performance of these algorithms across a range of datasets and application areas. It focuses on commonly used evaluation metrics including accuracy, precision, recall, and F1-score and examines the methodological rigor with which these metrics are applied. The review also discusses prevalent limitations in current evaluation practices and points to emerging strategies, such as hybrid models, transfer learning, and explainable artificial intelligence (XAI), as promising avenues for future investigation. The aim is to provide a thorough overview that supports the informed selection and optimization of classification algorithms for multi-class tasks in practical scenarios.

DOI: 10.48175/IJARSCT-26475

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, May 2025
Impact Factor: 7.67

II. METHODOLOGY

This review paper systematically analyzes and synthesizes recent research on the comparative performance of supervised machine learning classification algorithms on multi-class, labelled datasets. The methodology involved a structured literature search, application of clear inclusion and exclusion criteria, systematic data extraction, and qualitative synthesis of findings to identify trends and insights across diverse applications and datasets.

Literature Search Strategy:

The identification of pertinent studies was conducted through a structured search across a range of major academic databases and online repositories. These included, but were not limited to, platforms such as ScienceDirect, IEEE Xplore, SpringerLink, and various peer-reviewed journals and conference proceedings in the fields of computer science and data science. This approach ensured a comprehensive coverage of relevant literature from both well-established and emerging sources.

The primary keywords and search terms used in the literature search included combinations of general terms such as "classification algorithms", "comparative analysis" and "machine learning," along with the names of specific algorithms (e.g., "Decision Tree", "Support Vector Machine (SVM)", "Naïve Bayes (NB)", "K-Nearest Neighbour (KNN)", "Random Forest (RF)", "Gradient Boosting", "XGBoost", "Logistic Regression", and "Artificial Neural Networks (ANN)"). These were frequently combined with application-specific keywords such as "disease diagnosis," "sentiment analysis," and "data mining" to ensure comprehensive coverage of relevant studies across various domains.

Although the search strategy prioritized recent publications to reflect the current state of research, no strict date restrictions were initially imposed in order to capture foundational and influential studies. Nevertheless, the majority of sources included in this review were published within the last decade (approximately 2012–2023), aligning with contemporary research trends in the field.

Inclusion and Exclusion Criteria:

To ensure a rigorous and transparent study selection process, we applied the following predefined inclusion and exclusion criteria. Studies were included if they explicitly investigated supervised machine learning algorithms for classification tasks and offered a comparative analysis of two or more such algorithms on one or more datasets. Priority was given to papers presenting empirical evaluations of algorithm performance using standard metrics such as accuracy, precision, recall and F1-score and drawing on diverse datasets, including those related to disease diagnosis (e.g., heart disease, diabetes), sentiment analysis (e.g., Twitter data), and well-known benchmarks (such as the Iris dataset and resources from the UCI Machine Learning Repository). Crucially, each included study was required to describe its methodology in sufficient detail to ensure reproducibility and to report results clearly against the chosen performance criteria.

Conversely, we excluded studies that focused primarily on unsupervised learning techniques, regression tasks, or purely theoretical algorithm development lacking empirical validation. We also omitted research that fell outside the domains of general data classification, disease diagnosis, or sentiment analysis, unless it provided substantial comparative insights into the performance of the classification algorithms under consideration.

Data Extraction Process:

For each selected study, relevant information was systematically extracted, including:

- Classification algorithms: Recorded the specific supervised classification algorithms evaluated in each study.
- Datasets characteristics: Captured details of each dataset, including total sample size, number of features, and class distribution.
- Data preprocessing techniques: Noted any preprocessing steps applied (e.g., feature selection, normalization).
- Evaluation metrics: The metrics used to assess performance (e.g., accuracy, precision, recall, F1-score) as given in the referenced sources.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, May 2025
Impact Factor: 7.67

- Key findings and conclusions: Summarized comparative performance results, identifying which algorithms performed best under particular experimental conditions or on specific datasets.
- Influencing factors: Documented factors cited by authors as affecting performance, such as dataset size, feature-selection strategy, and hyperparameter optimization.

This information was organized to facilitate synthesis and comparison across the reviewed studies.

Synthesis and Analysis of Data:

The extracted data were synthesized using qualitative comparative analysis. Studies were organized according to the classification algorithms examined, the types of datasets employed, and the performance metrics reported. The findings were then compared to identify consistent trends, discrepancies, and factors that influence the effectiveness of various algorithms. The analysis primarily focused on several key areas. First, it assessed the relative performance of different classification algorithms, such as Decision Tree, SVM, Naïve Bayes, KNN, Random Forest, Gradient Boosting, XGBoost, Logistic Regression, and Artificial Neural Networks, across a range of datasets. For instance, while some studies identified Naïve Bayes as particularly effective, others emphasized the advantages of Random Forest or SVM, depending on the specific characteristics of the datasets. Second, the analysis examined the impact of dataset characteristics including size, dimensionality, class imbalance, and feature types on algorithm performance. Third, it evaluated the role of preprocessing techniques, such as feature selection, normalization, and the handling of missing values, in enhancing classification accuracy. Finally, the review considered domain-specific applications, highlighting algorithms that have demonstrated particular promise in areas such as disease diagnosis and sentiment analysis. This qualitative synthesis facilitated the identification of the general strengths and weaknesses of different classification algorithms and offered insights into their suitability for various applications and datasets.

Limitations of the Review

While this review aims to be comprehensive, it is subject to certain limitations. The selection of databases and keywords, although broad, may not have captured all relevant studies. The inclusion and exclusion criteria involved some degree of subjective judgment, which could introduce bias. Additionally, the qualitative synthesis, while providing valuable insights, does not constitute a quantitative meta-analysis due to the heterogeneity of datasets and evaluation metrics across studies. The review is primarily based on the information available in the selected sources, and the interpretation of findings depends on the methodologies and reporting standards of those studies.

By acknowledging these limitations, this methodology section provides a transparent overview of the process undertaken to conduct this review of machine learning classification algorithms.

III. DISCUSSION

This review synthesizes a wide range of studies that explore the application and performance of multi-class classification algorithms across diverse domains such as healthcare diagnostics, network security, social media sentiment analysis, and botanical classification. A consistent theme across the literature is that classifier performance is not determined solely by the algorithm itself but is significantly influenced by dataset characteristics, preprocessing methods, evaluation strategies, and the specific challenges of the application domain.

1. Synthesis of Key Findings

Algorithm Performance in Context: The reviewed studies demonstrate that no single algorithm consistently outperforms others across all datasets. Instead, performance is highly dependent on the context of use. Naïve Bayesian classifiers, for instance, are widely acknowledged for their simplicity, interpretability, and low computational cost. When the assumption of feature independence is satisfied, Naïve Bayes (NB) often performs competitively—even outperforming more complex models such as Decision Trees (DTs), Support Vector Machines (SVMs), and K-Nearest Neighbours (KNN) in certain cases [1]. For example, Gaussian Naïve Bayes achieved approximately 95% accuracy on

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, May 2025
Impact Factor: 7.67

the well-structured and balanced Iris dataset, demonstrating its effectiveness in scenarios with clean data and stable class distributions [9].

Support Vector Machines are particularly effective in handling high-dimensional feature spaces and complex decision boundaries [2], [4]. In a comparative study on a diabetes dataset, SVMs outperformed both Decision Trees and Naïve Bayes in terms of precision and accuracy [2]. Moreover, an advanced multi-class formulation of SVMs proposed in [4] eliminates the inefficiencies of one-vs-rest approaches, resulting in fewer kernel computations and improved classification accuracy. Despite these strengths, ensemble methods were observed to outperform SVMs in some critical healthcare applications, such as early heart disease diagnosis [3].

Decision Trees, including popular variants like J48, are appreciated for their interpretability and ease of use. However, they were shown to be less robust in some studies due to their tendency to overfit noisy or high-dimensional data [1], [8]. Scalability also becomes a limitation with larger datasets, though alternative implementations such as SPRINT and SLIQ have been proposed to address this challenge [1].

While conceptually simple, K-Nearest Neighbours has demonstrated strong performance in specialized domains. For instance, a KNN-based model used for Twitter sentiment analysis achieved an overall accuracy of approximately 86%, with a precision of 82% and recall of 81.5% [6]. However, the success of KNN is highly contingent upon effective feature extraction (e.g., N-gram models) and the appropriate choice of distance metrics [6].

Ensemble and Optimized Approaches: Beyond individual classifiers, ensemble methods and advanced optimization techniques have proven effective in improving prediction accuracy and handling complex data characteristics. Random Forest (RF), for example, has demonstrated notable robustness against overfitting and can model intricate feature interactions. An improved RF model presented in [7], which incorporates feature selection methods like Correlation-Based Feature Selection (CFS), Symmetrical Uncertainty, and Gain Ratio, along with instance filtering via resampling, significantly enhanced performance metrics such as F-measure, accuracy, and ROC sensitivity.

Gradient Boosting techniques, particularly XGBoost, also stood out for their performance. In a study focused on thyroid disease classification, an optimized XGBoost model achieved a remarkable 99% accuracy, emphasizing the importance of hyperparameter tuning and iterative boosting for uncovering subtle patterns in the data [10].

Artificial Neural Networks (ANNs), though less commonly featured in the reviewed literature, have shown competitive results in tasks such as thyroid disease prediction and anomaly detection in uranium datasets [2]. Meanwhile, Multinomial Logistic Regression (MLR) remains a valuable statistical tool for multi-class classification, offering interpretable results that highlight the impact of individual predictor variables. In a study on physical violence data, MLR effectively identified significant predictors and demonstrated suitability for domains where interpretability is paramount [5].

Evaluation Metrics and Methodological Rigor: The reviewed studies collectively emphasize the importance of employing a comprehensive set of evaluation metrics to fully capture model performance. While accuracy, precision, recall, and F1-score are most commonly reported, other measures such as the Kappa statistic, Mean Absolute Error (MAE), and Area Under the ROC Curve (AUC) are essential, particularly for imbalanced datasets or models deployed in critical decision-making contexts. Source [8] offers a detailed analysis of 16 metrics available in the WEKA platform, advocating for a multidimensional evaluation framework that assesses not only predictive accuracy but also robustness and reliability across data variations.

2. Critical Analysis of Methodologies

A recurring challenge in the literature involves striking a balance between model complexity and interpretability. Simpler models such as Naïve Bayes are widely valued for their transparency, ease of implementation, and low computational demands, making them particularly suitable for rapid deployment in environments where clarity and explainability are essential [1], [9]. In contrast, more complex algorithms like Support Vector Machines, ensemble techniques, and deep neural networks offer superior capacity to model non-linear relationships and high-dimensional data, often achieving higher predictive accuracy. However, these models frequently operate as “black boxes,” obscuring internal decision-making processes—a significant limitation in high-stakes fields like healthcare, where model interpretability is critical [2], [3], [10].

DOI: 10.48175/IJARSCT-26475

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, May 2025
Impact Factor: 7.67

Effective preprocessing techniques are repeatedly emphasized as pivotal in enhancing classifier performance. Studies have demonstrated that appropriate noise reduction, normalization, and feature selection significantly impact the robustness and accuracy of classifiers [7], [8]. For instance, the improved Random Forest classifier in [7] integrates correlation-based feature selection, symmetrical uncertainty, and gain ratio measures, along with resampling, to improve results on imbalanced datasets. Similarly, K-Nearest Neighbours was shown to perform well in sentiment analysis tasks only when robust feature engineering and distance metric optimization were applied [6].

Additionally, methodological rigor in evaluation practices is essential. Relying solely on accuracy can yield misleading interpretations of model effectiveness, particularly in the presence of class imbalance. Integrating a multidimensional suite of metrics including precision, recall, F1-score, Kappa statistic, and area under the ROC curve, provides a more nuanced and reliable understanding of classifier behaviour [8]. Source [8] offers an in-depth evaluation of 16 such metrics using WEKA, demonstrating how diverse metrics can expose both the strengths and weaknesses of classifiers under varying data conditions.

3. Practical Implications

The reviewed literature spans a wide array of domain-specific applications, each presenting distinct requirements and challenges for multi-class classification models. In the healthcare domain, for instance, accurate early diagnosis of conditions such as heart disease and thyroid disorders is critical due to the high consequences of misclassification. In these high-stakes settings, ensemble and optimized boosting algorithms offer notable advantages. Notably, Random Forest achieved superior accuracy in a heart disease prediction task [3], while an optimized XGBoost model attained nearly perfect classification accuracy (99%) on a thyroid disease dataset, outperforming traditional models through hyperparameter tuning and feature selection [10].

In contrast, social media analytics, such as sentiment classification on platforms like Twitter, require models that can handle unstructured and noisy textual data. In these cases, even relatively simple models like K-Nearest Neighbours (KNN) have shown competitive performance when enhanced with robust preprocessing techniques, such as N-gram feature extraction [6]. These results underscore the importance of tailoring model selection and preprocessing strategies to the specific characteristics of the domain and dataset.

Background literature cited within the reviewed studies also points to applications in educational performance prediction and network intrusion detection, where various classification techniques have been explored depending on data complexity and performance requirements [1].

Another key consideration in real-world deployment is the trade-off between computational efficiency and predictive accuracy. In resource-constrained or real-time environments, algorithms such as Naïve Bayes are often preferred due to their low memory and processing requirements [1]. However, in high-accuracy scenarios—particularly in fields like medical diagnostics—the additional computational overhead associated with potentially more accurate ensemble methods or deep learning architectures may be justified to ensure reliable and precise predictions [3], [10].

IV. LIMITATIONS AND FUTURE RESEARCH DIRECTIONS

The reviewed literature, while demonstrating substantial advancements in multi-class classification, also reveals a number of consistent limitations and highlights important directions for future work. A recurring observation across several studies is the strong dependency of classifier performance on specific dataset characteristics, such as the number of features, the size of the dataset, the level of noise, class imbalance, and the inherent separability of the classes [1], [2], [3], [8], [10]. For instance, in a study involving a Hepatitis C dataset, standard classifiers such as Naïve Bayes, BayesNet, J48, and Multilayer Perceptron failed to perform effectively, largely due to the dataset's complexity, class overlap, and insufficient preprocessing [8]. In contrast, high accuracy was observed on the Iris dataset, where a Gaussian Naïve Bayes classifier achieved strong results owing to the dataset's small size, balanced class distribution, and well-separated features [9]. Similarly, excellent performance was obtained on the Thyroid dataset; however, this was achieved using an optimized XGBoost model with carefully tuned hyperparameters and feature selection techniques, rather than with standard classifiers [10]. These findings highlight the critical influence of dataset characteristics and preprocessing strategies on classifier effectiveness across different domains.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, May 2025
Impact Factor: 7.67

This observation suggests a critical need for algorithms that are more adaptive and robust across a wide range of data conditions. One promising direction is the development of models capable of generalizing effectively across domains, possibly through the integration of transfer learning techniques or meta-learning frameworks that select optimal models based on dataset properties. Overfitting remains a pervasive issue, especially in models like decision trees [1], [3], as well as more complex architectures such as ensemble methods and neural networks [2], [3], [7], [10]. In particular, classifiers trained on noisy or informal datasets, such as those derived from social media, tend to suffer significant degradation in performance [6].

Addressing overfitting and noise requires further advances in regularization techniques tailored for multi-class problems and broader adoption of robust cross-validation strategies, such as the 10-fold cross-validation employed in several comparative analyses [7], [8]. Furthermore, improving data quality through advanced preprocessing is essential. Techniques such as noise reduction, outlier detection, feature engineering, and strategic feature selection have been shown to dramatically enhance performance, as evidenced by the improvements reported after applying such preprocessing methods in [7]. Conversely, the poor results reported in [8] when preprocessing was minimal emphasize the importance of these steps.

Another important limitation lies in the current evaluation frameworks. Most studies rely on a diverse set of metrics such as accuracy, precision, recall, and F1-score, yet these are often insufficient when dealing with imbalanced datasets or real-time requirements [1]–[3], [6]–[10]. There is a growing need for integrated evaluation frameworks that also account for computational efficiency, model interpretability, and real-world applicability. For example, while Naïve Bayes is fast and efficient [1], [9], it may lack the accuracy or adaptability of more complex models like ensembles, which can be computationally expensive [3], [10]. Similarly, models like Multinomial Logistic Regression are valued for their interpretability [5], an increasingly important factor in critical domains such as healthcare, whereas ensemble and neural network models often operate as black boxes.

Finally, hybrid and deep learning approaches offer considerable promise. The improved Random Forest classifier presented in [7] underscores the benefits of combining traditional models with enhanced techniques like feature selection and resampling. Future work should explore the design of hybrid models that integrate conventional algorithms (e.g., SVMs or Naïve Bayes) with ensemble or deep learning components to better handle high-dimensional or nonlinear data [1], [4], [10]. As deep learning continues to gain traction, particularly for handling large-scale and complex datasets, the issue of model transparency must be addressed. Future research should prioritize explainable AI (XAI) to make deep learning models more interpretable and trustworthy, especially in high-stakes fields like medicine, where understanding the rationale behind predictions is crucial, unlike in simpler models such as MLR [5].

V. CONCLUSION

In conclusion, the comparative analysis of multi-class classification algorithms reveals a dynamic interplay between algorithm selection, dataset characteristics, and application-specific needs. The literature consistently shows that there is no single best solution, with various investigations yielding diverse outcomes. While traditional methods like Naïve Bayesian have proven highly effective in certain studies and SVMs offer robust performance across applications, advanced ensemble methods like Random Forest and optimized boosting techniques such as XGBoost, which achieved 99% accuracy in thyroid disease classification, are pushing the boundaries of accuracy and robustness. The optimal classifier is determined not only by its inherent properties but critically by its alignment with the specific objectives and constraints of the application. Rigorous empirical validation using diverse metrics such as accuracy, precision, recall, and F1-score, comprehensive evaluation frameworks, and crucial preprocessing steps like feature selection are essential for advancing the state-of-the-art in multi-class classification. Future research should focus on integrating deep learning approaches to tackle complex data patterns, developing hybrid models that combine the strengths of different algorithms, and establishing robust, integrated evaluation strategies to address the diverse challenges inherent in these tasks.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, May 2025
Impact Factor: 7.67

REFERENCES

[1]. V. Sheth, U. Tripathi, and A. Sharma, "A Comparative analysis of Machine learning Algorithms for classification purpose," Procedia Computer Science, vol. 215, pp. 422–431, Jan. 2022, doi: 10.1016/j.procs.2022.12.044.

[2]. O. FY, A. JET, A. O, H. J. O, O. O, and A. J, "Supervised Machine Learning Algorithms: Classification and Comparison," International Journal of Computer Trends and Technology, vol. 48, no. 3, pp. 128–138, Jun. 2017, doi: 10.14445/22312803/ijctt-v48p126.

[3]. B. Akkaya, Istanbul University, N. Çolakoğlu, and T.C. Arel Üniversitesi, "Comparison of Multi-class Classification Algorithms on Early Diagnosis of Heart Diseases". ResearchGate Sept. 2019. [Online]. Available: https://www.researchgate.net/publication/338950098_Comparison_of_Multi-class_Classification_Algorithms_on_Early_Diagnosis_of_Heart_Diseases

[4]. J. Weston and C. Watkins, "Support vector machines for multi-class pattern recognition.," The European Symposium on Artificial Neural Networks, pp. 219–224, Jan. 1999, [Online]. Available: https://dblp.uni-trier.de/db/conf/esann/esann1999.html#WestonW99

[5]. M. El-Habil, "An Application on Multinomial Logistic Regression Model," Pakistan Journal of Statistics and Operation Research, vol. 8, no. 2, p. 271, Mar. 2012, doi: 10.18187/pjsor.v8i2.234.

[6]. S. Hota and S. Pathak, "KNN classifier based approach for multi-class sentiment analysis of twitter data," International Journal of Engineering & Technology, vol. 7, no. 3, p. 1372, Jul. 2018, doi: 10.14419/ijet.v7i3.12656.

[7]. A Chaudhary, S. Kolhe, and R. Kamal, "An improved random forest classifier for multi-class classification," Information Processing in Agriculture, vol. 3, no. 4, pp. 215–222, Sep. 2016, doi: 10.1016/j.inpa.2016.08.002.

[8]. Ž. Đ. Vujovic, "Classification Model Evaluation Metrics," International Journal of Advanced Computer Science and Applications, vol. 12, no. 6, Jan. 2021, doi: 10.14569/ijacsa.2021.0120670.

[9]. Z. Iqbal and M. Yadav, "Multiclass Classification with Iris Dataset using Gaussian Naïve Bayes," International Journal of Computer Science and Mobile Computing, vol. 9, no. 4, pp. 27–35, Apr. 2020.

[10]. M. Alnaggar, M. Handosa, T. Medhat, and M. Rashad, "Thyroid Disease Multi-class Classification based on Optimized Gradient Boosting Model", Egyptian Journal of Artificial Intelligence, vol. 2, no. 1, PP. 1-14, 2023.