JOURNAL
COMPUTER SCIENCES INSTITUTE

JCSI 38 (2026) 59-65
Received: 23 September 2025
Accepted: 12 October 2025

Comparative analysis of machine learning classifiers

Łukasz Krukowski*, Grzegorz Kozieł

Department of Computer Science, Lublin University of Technology, Nadbystrzycka 36B, 20-618 Lublin, Poland

Abstract

This study presents a comparative analysis of five machine learning classification algorithms: support vector machine (SVM), multilayer perceptron (MLP), classification and regression tree (CART), k-nearest neighbors algorithm (K-NN), and naive Bayes classifier (NB) across four datasets from various domains. Using nested cross-validation, the research evaluated classifier performance on Heart Disease, German Credit, Spambase, and Online Shoppers Purchasing Intention datasets. Results demonstrated that no single classifier consistently outperformed others across all datasets and selection should be based on dataset characteristics and application requirements. Dataset characteristics emerged as the primary factor influencing performance, with class imbalance proving particularly problematic. Training efficiency analysis revealed that simpler algorithms can maintain competitive performance with lower computational costs.

Keywords: machine learning; multilayer perceptron; classification and regression tree; naive Bayes classifier, support vector machine; k-nearest neighbors algorithm

*Corresponding author
Email address: s95460@pollub.edu.pl (Ł. Krukowski)

1. Introduction

Machine learning classification algorithms are the cornerstone of modern data science applications, from medical diagnosis and financial risk assessment to image recognition and natural language processing. Selecting the most appropriate classification algorithm has become a critical decision that directly impacts system performance, computational efficiency, and implementation success. This study conducts a systematic comparative analysis of five fundamental classification algorithms: support vector machine (SVM), multilayer perceptron (MLP), classification and regression tree (CART), k-nearest neighbors algorithm (K-NN) and Gaussian naive Bayes classifier (NB) across four datasets. Despite extensive research on individual algorithms, comprehensive empirical studies comparing different classifier performance across diverse datasets are lacking. Most comparative studies focus on algorithm-specific improvements or examine datasets only within specific domains. This study contributes to a broader understanding of machine learning classifier behavior.

2. Study objectives and motivation

This study quantifies the performance of classifiers across different datasets. The study examines how classifiers with different architectures perform on datasets from diverse domains with varying data characteristics, and analyzes the impact of dataset characteristics on classifier performance. Based on this objective, the following research hypothesis was formulated:

H1. No single classifier dominates performance across all datasets.

3. Literature review

The literature review encompassed comparative analyses and studies examining specific classifiers within particular domains. The review covers diverse topics. Medical studies include breast cancer diagnosis [1], diabetes [2], and chronic kidney disease analysis [3]. Cardiology research examined heart disease prediction [4][5], coronary artery disease prediction [6], and heart disease prevalence [7]. Other medical works include hypertension prediction [8] and medical decision-making studies [9]. Environmental studies covered rainfall forecasting [10], digital soil mapping [11], and land use change [12]. Other areas included fake news detection [13], network traffic classification [14], text classification analysis [15], image classification [16], drum sound classification [17], and noise detection and elimination in datasets [18]. Educational applications included student performance prediction [19] and research on education quality improvement [20]. Additional works covered multi-criteria ABC analysis [21] and bank soundness assessment [22].

Literature shows a lack of comprehensive comparative studies evaluating multiple classification techniques across diverse datasets and domains. Methodology varies in these studies, making comparisons between studies difficult. Literature review also revealed that certain classification techniques are widely used across fields. Multilayer perception (MLP) appeared most frequently in 12 articles, followed by classification and regression trees (DT) in 10 works. Naive Bayes (NB) classifier and support vector machines (SVM) each appeared in 9 studies. K-nearest neighbors algorithm (KNN) was the least common, used in 7 studies. Their prevalence suggests

fundamental importance and their comprehensive comparisons would improve understanding of them.

4. Methodology

The implementation utilized Python 3.11.9 with key libraries including scikit-learn 1.7.2, matplotlib 3.10.6, numpy 2.3.3, pandas 2.3.2. Experiment was concluded on a system equipped with system machine with AMD Ryzen R7 4700U (8-core/8-thread, 2.0 - 4.1 GHz) and 16GB of RAM (DDR4, 3200MT/s) running Windows 11 (24H2). Study covers dataset selection, data preprocessing and evaluation using nested cross-validation.

4.1. Dataset selection

Four datasets were selected for classification. The Heart Disease dataset [23] contains 303 patient records with 13 medical features collected from four medical institutions, designed for binary classification of heart disease presence balanced target with 53.9% negative cases. Features encompass demographic data, clinical measurements, and diagnostic test results, making it valuable for medical classification studies. The German Credit dataset [24] consists of 1,000 loan applications with 20 attributes describing personal, financial, and credit history details, used to classify applicants as good credit risks in 70 percent of cases and bad credit risks in 30 percent. This dataset reflects real-world financial decision-making scenarios with mixed categorical and numerical variables. The Spambase dataset [25] includes 4,601 email messages characterized by 57 features for spam detection, with 60.8% legitimate emails and 39.2% spam. Features represent word frequencies, character occurrences, and statistical measures of email content collected from actual spam reports and personal correspondence. The Online Shoppers Purchasing Intention (OSPI) dataset [26] contains 12,330 user sessions with 17 features predicting purchase conversion in e-commerce, where only 15.5% of sessions result in purchases. Features include page navigation patterns, time spent in different categories, bounce rates, and temporal factors like special days and weekends, with each session representing a unique user over one year. These datasets provide diverse classification challenges across healthcare, cybersecurity, finance, and e-commerce domains with varying class balance and feature types suitable for comparative algorithm evaluation.

4.2. Data preprocessing

The data processing follows a two-stage approach to prevent information leakage during cross-validation. Global preprocessing removes instances with missing values and converts targets to binary classification. The Heart Disease target is transformed from multi-level severity scores to a binary indicator of disease presence. Fold-specific processing applies different methods based on feature types, using only training data to maintain proper data isolation. Continuous features use StandardScaler normalization, non-ordinal categorical features employ OneHotEncoder, and ordinal features use OrdinalEncoder followed by StandardScaler. Binary categorical features use OrdinalEncoder without scaling. Features that required no processing remained unchanged. During cross-validation, unknown categories are mapped to 0 with OneHotEncoder or −1 with OrdinalEncoder to prevent model errors. Feature selection then retains 50% of features using SelectKBest with f_classif which is calculated based on the post-processing dimensions.

4.3. Evaluation design

This study employed nested cross-validation. The nested structure consists of two loops working together. The outer loop serves as an independent evaluation mechanism, completely isolating test data to ensure unbiased performance assessment. The inner loop handles hyperparameter optimization exclusively on training portions of the data. This separation prevents the common problem where the same data is used for both model tuning and performance evaluation, which typically inflates results since algorithms perform better on data, they were specifically tuned for [27]. Implemented nested cross-validation used a 5-fold outer loop and 5-fold inner loop configuration, with StratifiedKFold ensuring consistent class distributions across all folds. Hyperparameter optimization used GridSearchCV to test all parameter combinations, selecting the best performers based on accuracy. The same hyperparameter grid was used across all datasets to ensure fair comparison between algorithms. Measured metrics include accuracy, precision, recall, ROC-AUC and training time. Additionally, confusion matrices and learning curves were generated.

5. Results

The results presented are the averages obtained from five evaluations through cross-validation. All values have been rounded to three decimal places.

The results showed (Table 1) that MLP had three times the highest accuracy score (Heart Disease: 0.838, Spambase: 0.929, OSPI: 0.898) while SVM had the top score once (German Credit: 0.770). NB had twice the lowest score (Spambase: 0.905, OSPI: 0.847). DT and KNN had the lowest score once each (German Credit: 0.727, Heart Disease: 0.801). However, the differences were small. Classifiers had comparable performance with average difference between the highest and the lowest accuracy score being 0.041. The precision and sensitivity results showed greater differences, but most classifiers still maintained similar results. The best overall results classifiers achieved for the Heart Disease and Spambase datasets. The lowest results were for the German Credit dataset, particularly lacking in precision and sensitivity scores. The OSPI dataset, despite good accuracy scores, fell behind with the other metrics.

Table 1: Detailed results

Dataset	Classifier	Accuracy	Precision	Recall
Heart Disease	SVM	0.835 ±0.069	0.853 ±0.074	0.773 ±0.098
	MLP	0.838 ±0.037	0.846 ±0.039	0.796 ±0.067
	DT	0.815 ±0.072	0.821 ±0.093	0.773 ±0.087
	KNN	0.801 ±0.057	0.818 ±0.090	0.744 ±0.078
	NB	0.825 ±0.083	0.815 ±0.101	0.809 ±0.077
German Credit	SVM	0.770 ±0.014	0.659 ±0.022	0.480 ±0.052
	MLP	0.751 ±0.017	0.621 ±0.047	0.473 ±0.082
	DT	0.727 ±0.035	0.560 ±0.077	0.443 ±0.050
	KNN	0.752 ±0.020	0.643 ±0.080	0.413 ±0.073
	NB	0.750 ±0.016	0.627 ±0.038	0.410 ±0.096
Spambase	SVM	0.927 ±0.006	0.919 ±0.014	0.894 ±0.017
	MLP	0.929 ±0.002	0.926 ±0.015	0.891 ±0.017
	DT	0.923 ±0.004	0.913 ±0.006	0.891 ±0.015
	KNN	0.905 ±0.008	0.924 ±0.008	0.828 ±0.018
	NB	0.897 ±0.005	0.861 ±0.015	0.883 ±0.012
OSPI	SVM	0.894 ±0.005	0.715 ±0.020	0.523 ±0.032
	MLP	0.898 ±0.006	0.715 ±0.027	0.568 ±0.040
	DT	0.893 ±0.004	0.691 ±0.026	0.562 ±0.033
	KNN	0.882 ±0.004	0.727 ±0.015	0.380 ±0.032
	NB	0.847 ±0.004	0.506 ±0.013	0.450 ±0.029

The AUC results (Figure 1) were good regardless of the dataset and were also similar for all classifiers. The average AUC score was 0.893 for Heart Disease, 0.766 for German Credit, 0.958 for Spambase, and 0.871 for OSPI.

The training times were consistently the shortest for DT, KNN, and NB (Figure 2). As the dataset size increased, the training times for SVM and MLP increased significantly.

The confusion matrices confirm the obtained results (Figures for 3 to 6). Heart Disease and Spambase had the highest rate of correct classifications with a small rate of incorrect classifications. In the other datasets, the rate of incorrect classifications was quite high compared to correct classifications of the positive class, and even exceeded it in the German Credit dataset.

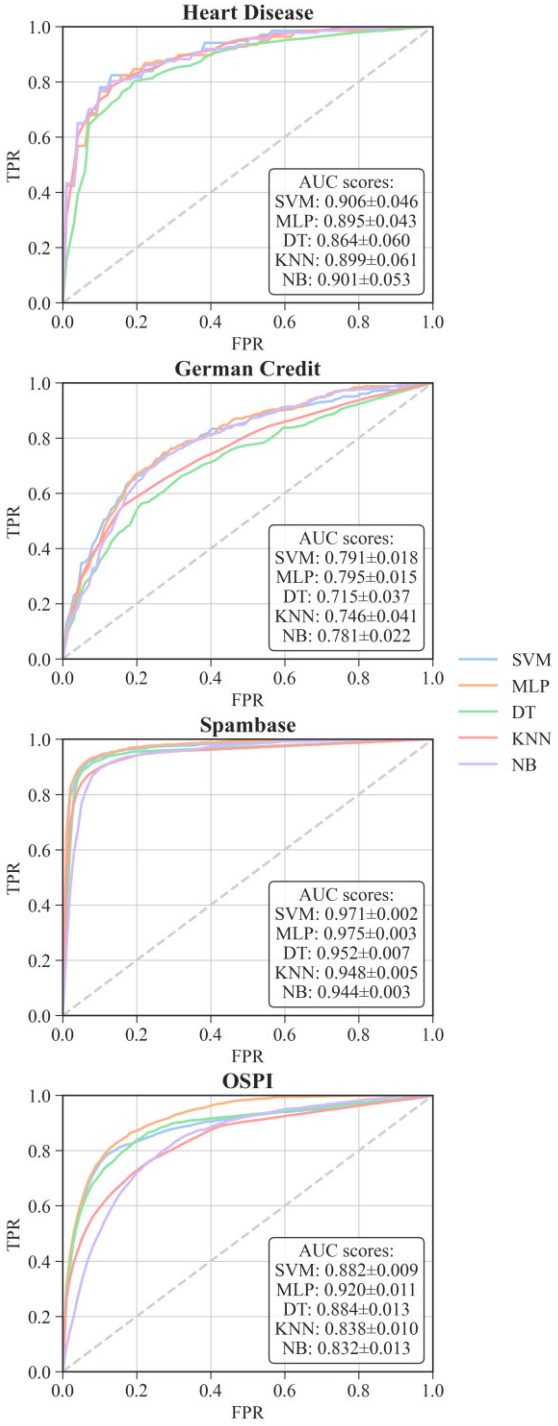

Figure 1: ROC curves with AUC scores.

The learning curves show different behaviors (Figures from 7 to 10). In the Heart Disease dataset, all classifiers quickly learned to generalize well. The German Credit dataset showed that increasing the amount of data did not significantly increase performance, quickly reaching a limit. In the Spambase dataset, all classifiers showed consistent improvement in performance as more data became available. The OSPI dataset showed similar patterns to the German Credit dataset.

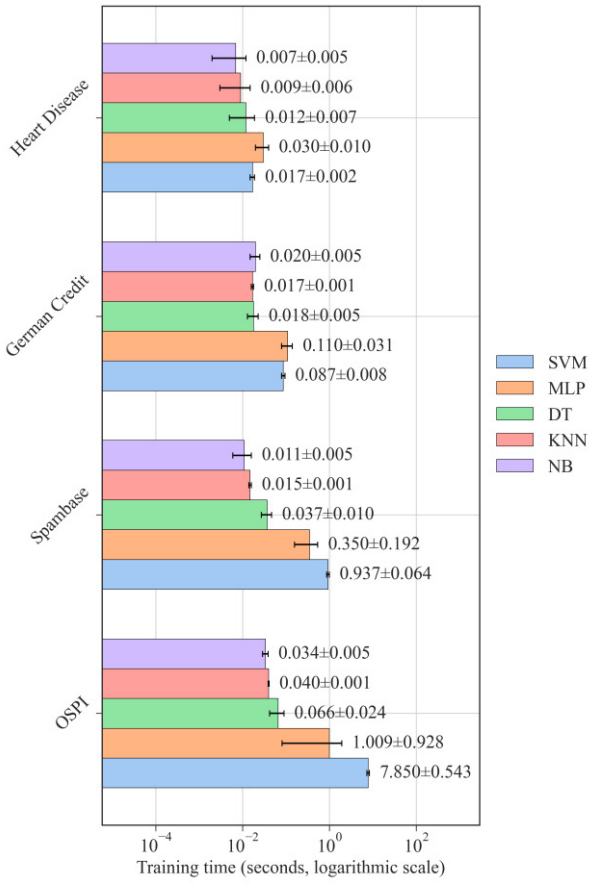

Figure 2: Training times.

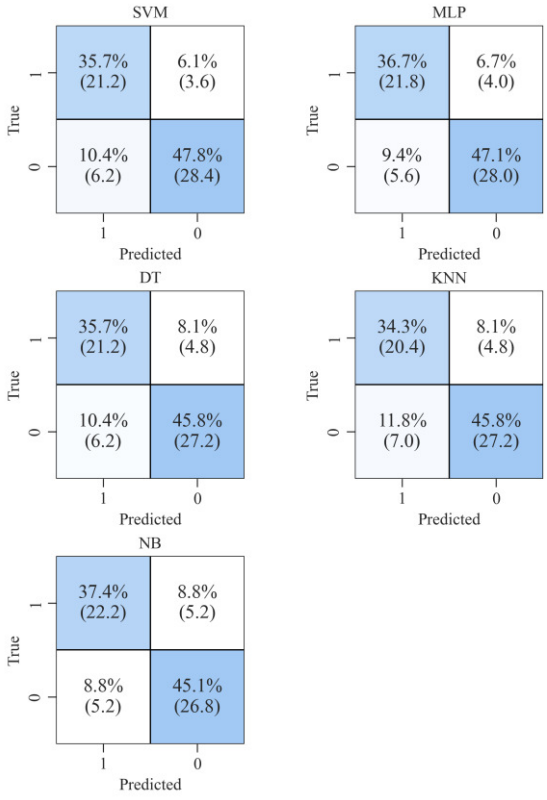

Figure 3: Confusion matrices Heart Disease dataset.

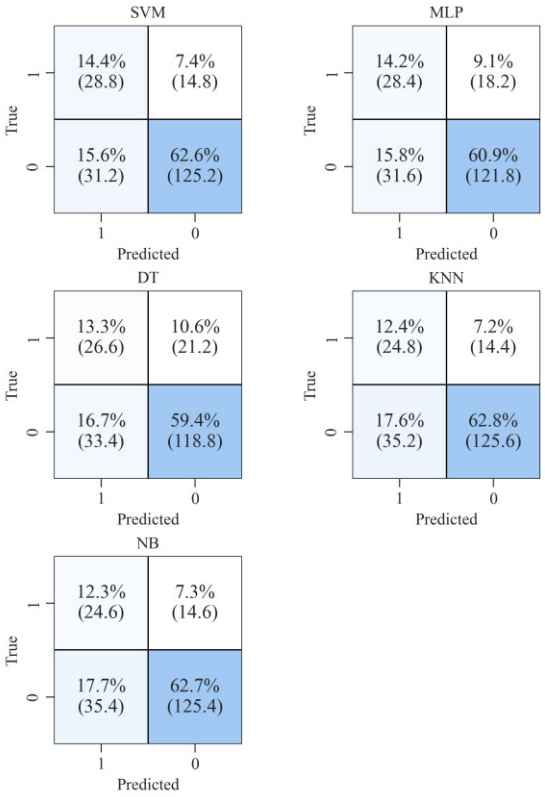

Figure 4: Confusion matrices German Credit dataset.

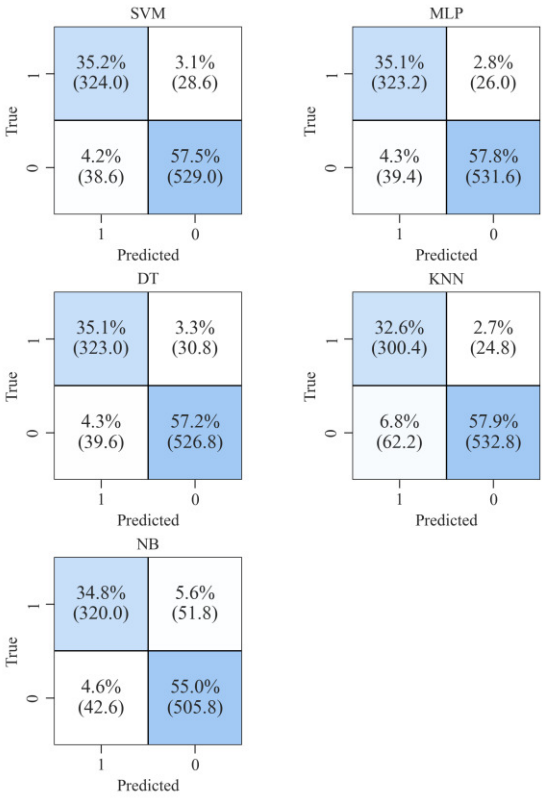

Figure 5: Confusion matrices for Spambase.

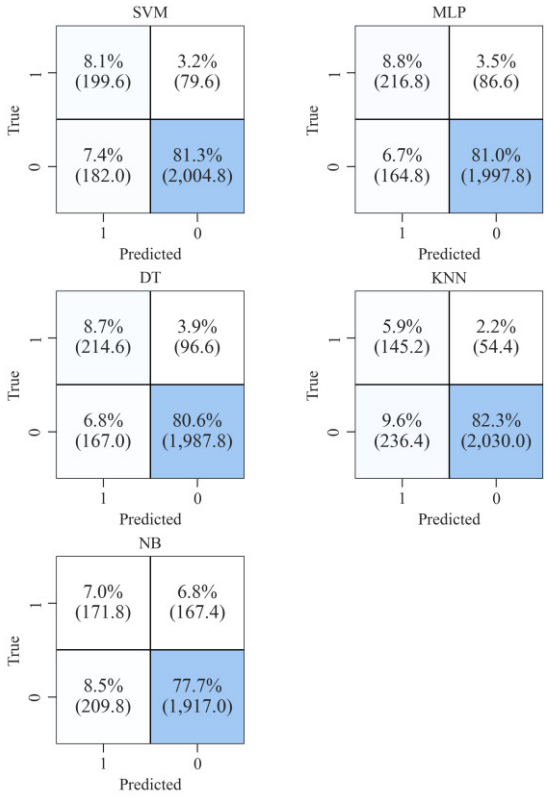

Figure 6: Confusion matrices for OSPI.

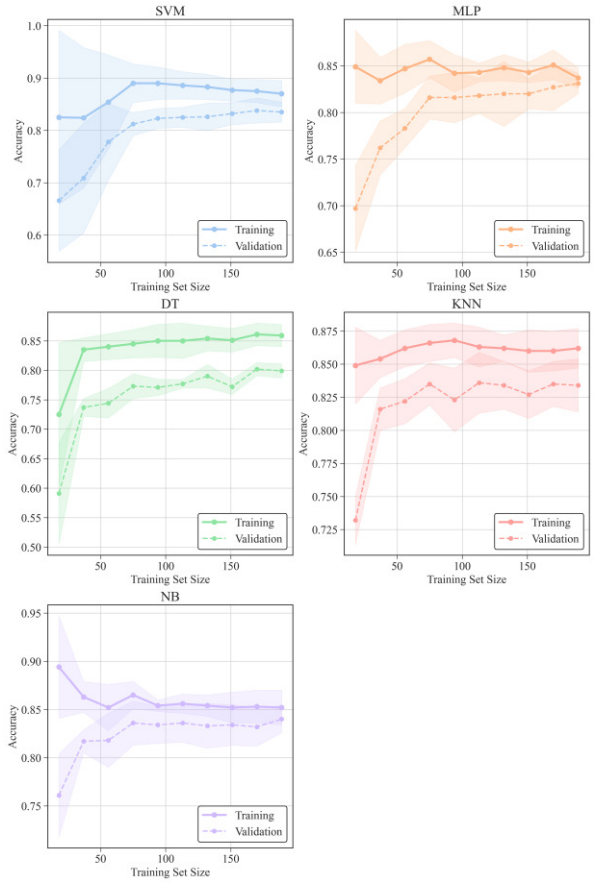

Figure 7: Learning curves for Heart Disease.

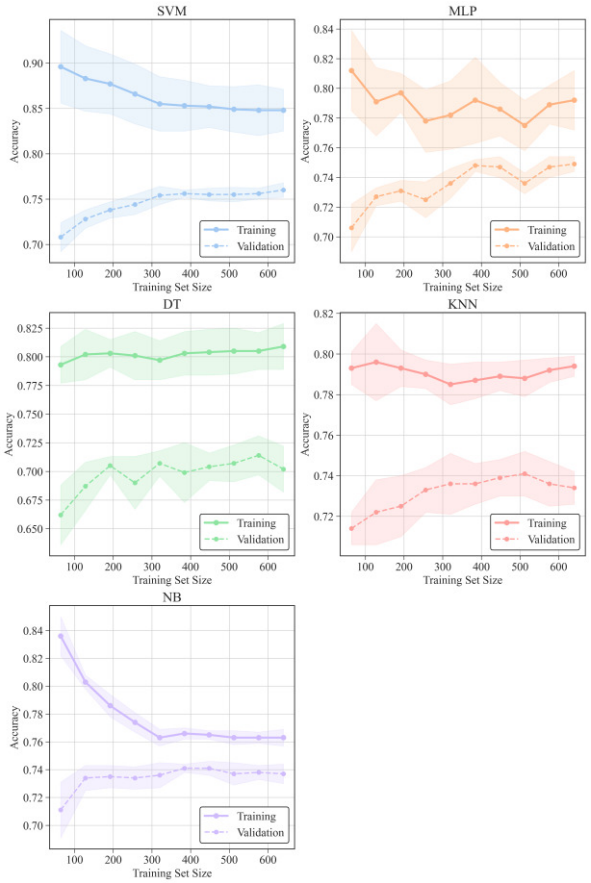

Figure 8: Learning curves for Geman Credit.

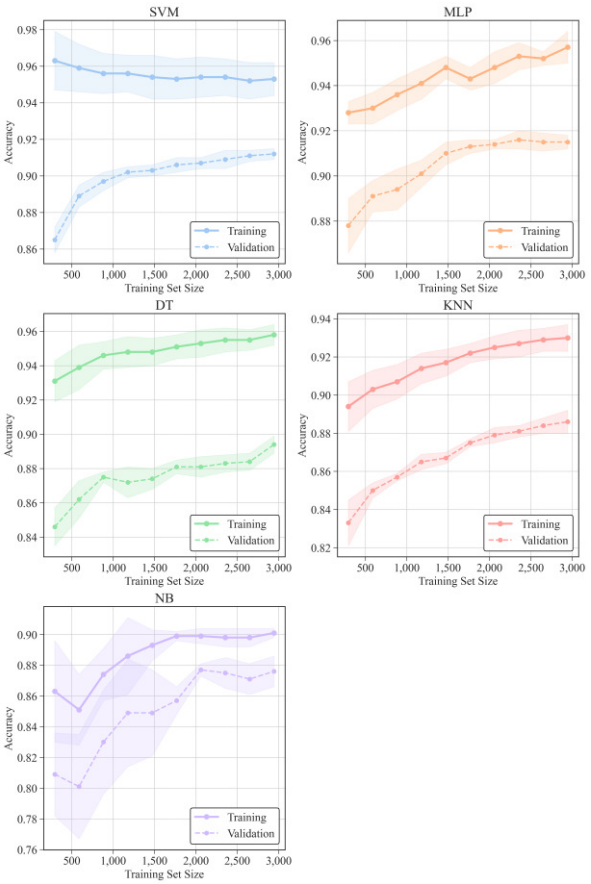

Figure 9: Learning curves for Spambase.

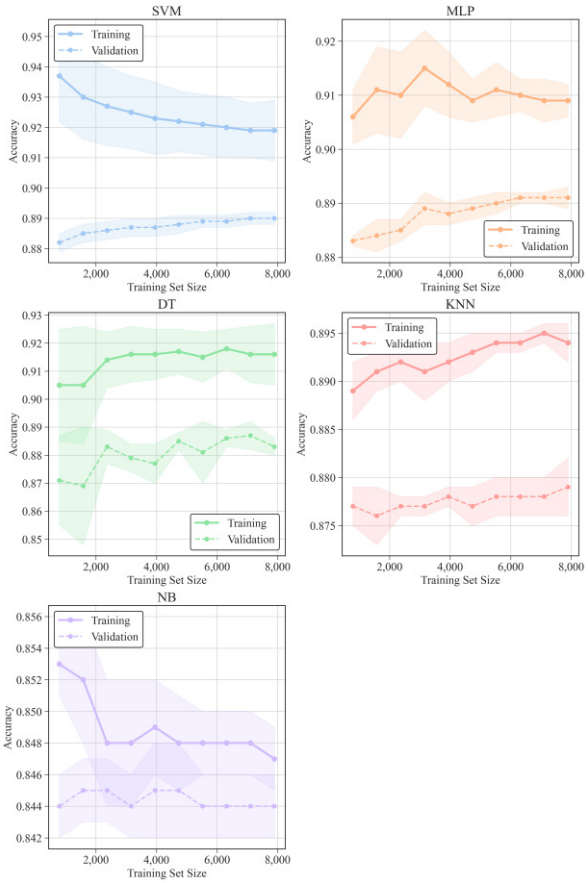

Figure 10: Learning curves OSPI.

6. Discussion

The results demonstrate that no single classifier consistently outperformed others across all datasets. However, the performance differences were relatively small compared to other classifiers, which frequently offered competitive results and sometimes performed better in specific metrics (Table 1). On the Heart Disease dataset, while MLP achieved the highest accuracy, SVM demonstrated superior precision (0.853) and NB achieved the highest recall (0.809). Similarly, on the Online Shopping dataset, KNN achieved the highest precision (0.727) despite having lower overall accuracy than MLP. These metric-specific advantages highlight the importance of considering individual performance requirements, as different applications may prioritize precision over recall or vice versa. This pattern suggests that classifier selection should be context-dependent rather than based on general superiority assumptions.

Despite varying accuracy and precision-recall performance, ROC-AUC scores remained consistently high across all datasets, indicating strong class discrimination capabilities for all classifiers (Figure 1) even on challenging datasets like German Credit. It suggests that poor precision-recall performance may stem from suboptimal threshold selection rather than fundamental classification inability and results could be improved.

Dataset properties emerged as the most critical factor influencing classifier performance. Class imbalance proved particularly problematic, as evidenced in the German Credit and Online Shopping datasets, where all classifiers struggled with positive class prediction, often with higher false positive rates than true positives.

The learning curve analysis further support class imbalanced impact on performance. On the Spambase dataset, most classifiers showed continuous improvement with increased data while on German Credit and Online Shopping datasets, learning curves plateaued early, which might be caused by class imbalance rather than insufficient data

The standard deviation analysis revealed that larger datasets promoted more stable performance, with variability decreasing as dataset size increased. This stability improvement suggests that an adequate sample size is crucial for reliable classifier evaluation and deployment.

Training time scores revealed that longer training does not always guarantee better results. DT, KNN and NB consistently required minimal training time across all datasets (0.007-0.066 seconds) while maintaining competitive performance levels (Figure 2). This efficiency advantage became particularly pronounced on larger datasets, where SVM training time increased dramatically to 7.85 seconds on the Online Shopping dataset, compared to 0.040 seconds for KNN while offering similar performance.

7. Conclusion

This comprehensive evaluation of five machine learning classifiers across four diverse datasets confirms the hypothesis that no single classifier performs better across all datasets. The optimal classifier selection depends on dataset characteristics, computational constraints, and application requirements.

References

[1] D. A. Omondiagbe, S. Veeramani, A. S. Sidhu, Machine Learning Classification Techniques for Breast Cancer Diagnosis, IOP Conf Ser Mater Sci Eng 495 (2019) 012033, https://doi.org/10.1088/1757-899X/495/1/012033.

[2] R. M. Rahman, F. Afroz, Comparison of Various Classification Techniques Using Different Data Mining Tools for Diabetes Diagnosis, Journal of Software Engineering and Applications 6 (2013) 85–97, https://doi.org/10.4236/jsea.2013.63013.

[3] V. Kunwar, K. Chandel, A. S. Sabitha, A. Bansal, Chronic Kidney Disease analysis using data mining classification techniques, Proceedings of the 2016 6th International Conference - Cloud System and Big Data Engineering (2016) 300–305, https://doi.org/10.1109/CONFLUENCE.2016.7508132.

[4] C. S. Dangare, S. S. Apte, Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques, Int J Comput Appl 47 (2012) 44–48, https://doi.org/10.5120/7228-0076.

[5] C. B. C. Latha, S. C. Jeeva, Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques, Inform Med Unlocked 16 (2019) 100203, https://doi.org/10.1016/j.imu.2019.100203.

[6] I. Kurt, M. Ture, A. T. Kurum, Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease, Expert Syst Appl 34 (2008) 366–374, https://doi.org/10.1016/j.eswa.2006.09.004.

[7] D. Khanna, R. Sahu, V. Baths, B. Deshpande, Comparative Study of Classification Techniques (SVM, Logistic Regression and Neural Networks) to Predict the Prevalence of Heart Disease, Int J Mach Learn Comput 5 (2015) 414–419, https://doi.org/10.7763/ijmlc.2015.v5.544.

[8] M. Ture, I. Kurt, A. Turhan Kurum, K. Ozdamar, Comparing classification techniques for predicting essential hypertension, Expert Syst Appl 29 (2005) 583–588, https://doi.org/10.1016/j.eswa.2005.04.014.

[9] P. R. Harper, A review and comparison of classification algorithms for medical decision making, Health Policy 71 (2005) 315–331, https://doi.org/10.1016/j.healthpol.2004.05.002.

[10] D. Gupta, U. Ghose, A comparative study of classification algorithms for forecasting rainfall, In 2015 4th International Conference on Reliability, Infocom Technologies and Optimization: Trends and Future Directions (2015) 1–6, https://doi.org/10.1109/ICRITO.2015.7359273.

[11] B. Heung, H. C. Ho, J. Zhang, A. Knudby, C. E. Bulmer, M. G. Schmidt, An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping, Geoderma 265 (2016) 62–77, https://doi.org/10.1016/j.geoderma.2015.11.014.

[12] P. K. Srivastava, D. Han, M. A. Rico-Ramirez, M. Bray, T. Islam, Selection of classification techniques for land use/land cover change investigation, Advances in Space Research 50 (2012) 1250–1265, https://doi.org/10.1016/j.asr.2012.06.032.

[13] M. Park, S. Chai, Constructing a User-Centered Fake News Detection Model by Using Classification Algorithms in Machine Learning Techniques, IEEE Access 11 (2023) 71517–71527, https://doi.org/10.1109/ACCESS.2023.3294613.

[14] M. Shafiq, X. Yu, A. A. Laghari, L. Yao, N. K. Karn, F. Abdessamia, Network Traffic Classification techniques and comparative analysis using Machine Learning algorithms, In 2016 2nd IEEE International Conference on Computer and Communications (2016) 2451–2455, https://doi.org/10.1109/COMPCOMM.2016.7925139.

[15] M. Thangaraj, M. Sivakami, Text classification techniques: A literature review, Interdisciplinary Journal of Information, Knowledge, and Management 13 (2018) 117–135, https://doi.org/10.28945/4066.

[16] S. Shakya, Analysis of Artificial Intelligence based Image Classification Techniques, Journal of Innovative Image Processing 2 (2020) 44–54, https://doi.org/10.36548/jiip.2020.1.005.

[17] P. Herrera, A. Yeterian, F. Gouyon, Automatic Classification of Drum Sounds: A Comparison of Feature Selection Methods and Classification Techniques, In: C. Anagnostopoulou, M. Ferrand, A. Smaill, (eds) Music and Artificial Intelligence. ICMAI 2002. Lecture Notes in Computer Science, Springer 2445 (2002) 69–80, https://doi.org/10.1007/3-540-45722-4_8.

[18] A. L. B. Miranda, L. P. F. Garcia, A. C. P. L. F. Carvalho, A. C. Lorena, Use of Classification Algorithms in Noise Detection and Elimination, In: E. Corchado, X. Wu, E. Oja, Á. Herrero, B. Baruque, (eds) Hybrid Artificial Intelligence Systems. HAIS 2009. Lecture Notes in Computer Science, Springer 5572 (2009) 417–424, https://doi.org/10.1007/978-3-642-02319-4_50.

[19] M. Mayilvaganan, D. Kalpanadevi, Comparison of classification techniques for predicting the performance of students academic environment, In 2014 International Conference on Communication and Network Technologies (2014) 113–118, https://doi.org/10.1109/CNT.2014.7062736.

[20] K. Bunkar, U. K. Singh, B. Pandya, R. Bunkar, Data mining: Prediction for performance improvement of graduate students using classification, In 2012 Ninth International Conference on Wireless and Optical Communications Networks (2012) 1–5, https://doi.org/10.1109/WOCN.2012.6335530.

[21] M. C. Yu, Multi-criteria ABC analysis using artificial-intelligence-based classification techniques, Expert Syst Appl 38 (2011) 3416–3421, https://doi.org/10.1016/j.eswa.2010.08.127.

[22] C. Ioannidis, F. Pasiouras, C. Zopounidis, Assessing bank soundness with classification techniques, Omega 38 (2010) 345–357, https://doi.org/10.1016/j.omega.2009.10.009.

[23] Heart Disease - UCI Machine Learning Repository, https://archive.ics.uci.edu/dataset/45/heart+disease [05.08.2025].

[24] Statlog (German Credit Data) - UCI Machine Learning Repository, https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data [06.08.2025].

[25] Spambase - UCI Machine Learning Repository, https://archive.ics.uci.edu/dataset/94/spambase [06.08.2025].

[26] Online Shoppers Purchasing Intention Dataset - UCI Machine Learning Repository, https://archive.ics.uci.edu/dataset/468/online+shoppers+purchasing+intention+dataset [11.08.2025].

[27] G. C. Cawley, N. L.C. Talbot, On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation, The Journal of Machine Learning Research 11 (2010) 2079–2107, https://doi.org/10.5555/1756006.1859921.