DOI: 10.65322/j.srd.05192026.68

COMPARATIVE EVALUATION OF SIX MACHINE LEARNING CLASSIFIERS ON THREE PUBLIC DATASETS USING REPEATED STRATIFIED CROSS-VALIDATION

Sayidov N.A., *PhD, dotsent,*
Ergasheva Z.O., *4-bosqich talabasi,*
E-mail: sayidovnozimjon@gmail.com
Namangan davlat texnika universiteti

Abstract: *To compare the predictive performance and training cost of six standard supervised classifiers on three small public benchmark datasets without claiming a new algorithm. Methods: Logistic regression, k-nearest neighbours (k-NN), radial-basis-function support vector machine (RBF-SVM), random forest, gradient boosting and a two-hidden-layer multilayer perceptron (MLP) were assessed in the source study using three repeats of stratified 10-fold cross-validation. Results: Across the three datasets, mean accuracy was 97.94% for SVM, 97.93% for logistic regression and MLP, 97.37% for random forest, 97.20% for k-NN and 95.59% for gradient boosting. On Digits, the reported mean fit times were 0.051 s for SVM and 6.726 s for gradient boosting; these are descriptive hardware-dependent observations. Conclusion: Untuned model rankings depend on the dataset, and accuracy should be evaluated alongside uncertainty and computational resources. Raw fold results and executable code were unavailable for independent verification.*

Annotatsiya: *Oltita standart klassifikatorning uchta ochiq ma'lumotlar to'plamidagi aniqligi va o'qitish xarajatini qiyoslash. Usullar: logistik regressiya, k-NN, RBF-SVM, tasodifiy o'rmon, gradient boosting va ikki yashirin qatlamli MLP uch marta takrorlangan 10 qismli stratifikatsiyalangan kross-validatsiya orqali baholangan. Natijalar: uch to'plam bo'yicha o'rtacha aniqlik SVM uchun 97,94%, logistik regressiya va MLP uchun 97,93%, tasodifiy o'rmon uchun 97,37%, k-NN uchun 97,20% va gradient boosting uchun 95,59% bo'lgan. Digits to'plamida SVM va gradient boosting o'qitish vaqti mos ravishda 0,051 va 6,726 soniya deb qayd etilgan. Xulosa: sozlanmagan modellarning reytingi ma'lumotlarga bog'liq; aniqlikni tarqoqlik va hisoblash xarajati bilan birga baholash zarur. Xom natijalar hamda bajariladigan kod mavjud emasligi mustaqil qayta tekshirishni cheklaydi.*

Аннотация: *Цель: сопоставить точность и вычислительные затраты шести стандартных классификаторов на трёх открытых наборах данных. Методы: логистическая регрессия, k-NN, RBF-SVM, случайный лес, градиентный бустинг и двухслойный MLP оценивались с помощью стратифицированной 10-кратной кросс-валидации с тремя повторами. Результаты: средняя точность по трём наборам составляла 97,94% для SVM, 97,93% для логистической регрессии и MLP, 97,37% для случайного леса, 97,20% для k-NN и 95,59% для градиентного бустинга. Для Digits время обучения SVM и градиентного бустинга составляло соответственно 0,051 и 6,726 с. Вывод: ранжирование ненастроенных моделей зависит от данных; важны вариабельность результатов и вычислительные затраты. Отсутствие исходных результатов по фолдам и исполняемого кода ограничивает независимую проверку.*

DOI: 10.65322/j.srd.05192026.68

Keywords: *classification; repeated stratified cross-validation; benchmark datasets; model comparison; training time; reproducibility.*

Kalit so‘zlar: *tasniflash; takroriy stratifikatsiyalangan kross-validatsiya; etalon ma’lumotlar to‘plamlari; modellarni qiyoslash; o‘qitish vaqti; natijalarni qayta hosil qilish imkoniyati.*

Ключевые слова: *классификация; повторная стратифицированная кросс-валидация; эталонные наборы данных; сравнение моделей; время обучения; воспроизводимость результатов.*

Introduction. Machine-learning practitioners often need a defensible baseline before deploying a more computationally intensive system. Predictive performance depends on the geometry, scale, dimensionality and class structure of the available data; consequently, an isolated accuracy figure does not establish universal algorithmic superiority. Statistical learning theory and comparative benchmarking motivate evaluation on unseen observations and explicit reporting of uncertainty [1, 2].

This study revisits an existing controlled experiment rather than proposing an original classifier. It compares six familiar algorithm families on Breast Cancer Wisconsin (Diagnostic), Wine and Digits as distributed through scikit-learn [3]. These educational benchmarks provide transparent settings for illustrating model-selection trade-offs but are insufficient to establish performance on industrial or clinical operational data. Earlier large-scale comparisons show that model behaviour on tabular data is heterogeneous and highly dependent on protocol and tuning [4].

The contribution is a clearly delimited descriptive synthesis of the authors’ reported repeated-cross-validation measurements, with explicit attention to performance gaps, fold variability, compute time and inferential limitations. The existing dataset–algorithm matrix is retained without fabricating additional training runs, tuned-model scores or diagnostic measures.

Research questions

RQ1. How do the reported accuracy and macro-F1 values of six untuned classifiers differ across three heterogeneous public datasets?

RQ2. What can reasonably be inferred about the observed performance differences from the reported variability and multiplicity-adjusted comparisons?

RQ3. How does reported training cost–and, on Digits, training-set size–inform baseline model selection?

Scope and novelty

The experiment is a methodological baseline and its novelty should not be overstated. A stronger novel contribution would require pre-specified hyperparameter search nested within outer validation, materially broader datasets, and ideally application-specific prospective or external data. Nested validation is particularly important when model selection itself is estimated on the same sample [5]. The present manuscript does not represent such an additional experiment.

Methods. *2.1. Data and models.* Three publicly accessible datasets were used in the original experiment. They differ in sample count, number of predictors and class count (Table 1). Breast Cancer Wisconsin contains diagnostic measurements and is analysed here solely as an educational benchmark; no clinical diagnostic utility is claimed.

DOI: 10.65322/j.srd.05192026.68

1-jadval

Datasets and task characteristics

Dataset	Cases	Features	Classes	Prediction target
Breast Cancer Wisconsin	569	30	2	Benign / malignant
Wine	178	13	3	Wine class
Digits	1,797	64	10	Handwritten digit

For scale-sensitive models–logistic regression, k-NN, RBF-SVM and MLP–standardisation was reportedly fitted to each training fold only and then applied to that fold's test portion. Such train-only preprocessing is necessary to avoid leakage. Tree-based methods used the original feature scales.

2-jadval

Classifiers and specified settings in the source experiment

Model	Configuration
Logistic regression	L2, C=1; max_iter=2,000; scaled
k-NN	5 neighbours; Euclidean distance; scaled
RBF-SVM	C=1; gamma=scale; scaled
Random forest	300 trees; random_state=42
Gradient boosting	100 estimators; max_depth=3; learning_rate=0.1; seed=42
MLP (64–32)	ReLU, Adam, 64 and 32 hidden units; max_iter=1,000; scaled; seed=42

2.2. Evaluation and analysis protocol

The source study reports 10-fold stratified cross-validation repeated three times (30 held-out folds for each of 18 model–dataset combinations), using a fixed seed of 42. Accuracy is the share of correctly predicted observations. Macro-F1 is the arithmetic mean of class-wise F1 values and assigns equal weight to each class. The published summaries are mean accuracy ± standard deviation across 30 fold scores; macro-F1 is provided as a mean only. The fold scores should not be treated as 30 independent samples.

The authors additionally reported corrected resampled t-tests against each dataset's top classifier. They conducted 15 pairwise comparisons and applied a Bonferroni threshold of 0.05/15 ≈ 0.0033. Without fold-level predictions and exact test implementation, the reported p-values cannot be independently audited; they are presented as source-reported values, not newly computed estimates [6].

Learning curves were reported on Digits using 5-fold stratified validation over increasing training subsets. Fit time was reported as the mean wall-clock duration per fit on a two-core CPU. Neither repeated benchmark timings with dispersion nor processor model and software environment manifest was supplied. A figure caption in the source mentions scikit-learn 1.9.1, but the complete environment has not been validated.

2.3. Reproducibility and provenance

DOI: 10.65322/j.srd.05192026.68

Data identifiers: sklearn.datasets. load_breast_cancer, load_wine, and load_digits (scikit-learn). The source lists fold strategy, primary model settings and the random seed; however, no authoritative executable script, repository URL, dependency lockfile, per-fold output, hardware specification, or verified code hash accompanies the manuscript. Thus the protocol is partly documented but numerical reproducibility has not been established.

To enable independent replication, the authors should deposit a version-controlled script and README recording Python, numpy, scipy and scikit-learn versions, CPU details, thread counts, exact repeated-split object, random states, pipelines, timing procedure, complete fold predictions and raw scores. No repository link is supplied here because one has not been verified.

2.4. Interpretation constraints

An arithmetic mean of three dataset-level accuracies weights each dataset equally, not each patient or sample. Comparison of these three teaching datasets is descriptive; it is not a population estimate for real-world tabular problems. Friedman–Nemenyi methods are established for repeated comparisons over datasets [7], but three heterogeneous datasets offer too few independent tasks for strong across-dataset rank inference.

Results. The following measurements are transcribed from the supplied source manuscript and have not been independently rerun. Each accuracy cell is a reported mean ± fold standard deviation in percentage points (pp).

3-jadval

Reported accuracy, macro-F1 and fit time (six classifiers × three datasets)

Dataset	Classifier	Accuracy (%)	Macro-F1 (%)	Fit (s)
Breast Cancer	Logistic regression	97.78 ± 1.69	97.60	0.004
	k-NN	96.90 ± 1.79	96.62	0.002
	RBF-SVM	97.54 ± 1.90	97.36	0.005
	Random forest	95.96 ± 2.83	95.64	0.525
	Gradient boosting	95.96 ± 2.57	95.64	0.370
	MLP (64–32)	97.66 ± 1.72	97.47	0.273
Wine	Logistic regression	98.87 ± 2.27	98.84	0.004
	k-NN	96.99 ± 4.33	97.06	0.002
	RBF-SVM	97.94 ± 3.06	97.89	0.002
	Random forest	98.30 ± 2.97	98.34	0.469
	Gradient boosting	94.20 ± 5.50	94.34	0.295
	MLP (64–32)	98.49 ± 2.51	98.44	0.121
Digits	Logistic regression	97.14 ± 0.98	97.15	0.337
	k-NN	97.72 ± 0.95	97.72	0.003
	RBF-SVM	98.35 ± 0.87	98.34	0.051
	Random forest	97.85 ± 0.93	97.84	0.767
	Gradient boosting	96.61 ± 1.08	96.61	6.726
	MLP (64–32)	97.64 ± 0.89	97.63	0.704

DOI: 10.65322/j.srd.05192026.68

The highest reported accuracy was obtained by logistic regression on Breast Cancer (97.78%) and Wine (98.87%), and by RBF-SVM on Digits (98.35%). Mean accuracy across datasets was 97.94% (SVM), 97.93% (logistic regression), 97.93% (MLP), 97.37% (random forest), 97.20% (k-NN), and 95.59% (gradient boosting), rounded to two decimals. These are equal-weight dataset averages rather than estimates of deployment performance.

Across all 18 rows, the fold standard deviations range from 0.87 pp (SVM on Digits) to 5.50 pp (gradient boosting on Wine). For the best three models within each individual dataset, the relevant observed standard deviations fall within 0.87–2.97 pp; the 0.9–3.1 pp interval in the original abstract was rounded differently and is replaced here with the table-consistent precision.

Analysis of the results

The local separation between the leading classifiers is modest relative to across-fold score variation. On Breast Cancer, the best three results are 97.78%, 97.66% and 97.54%, a range of 0.24 pp; on Wine, the leading values are 98.87%, 98.49% and 98.30%, a range of 0.57 pp; on Digits, they are 98.35%, 97.85% and 97.72%, a range of 0.63 pp. Fold-level standard deviation is not, by itself, a confidence interval for paired differences, so overlap should not be used as a substitute for formal inference.

4-jadval

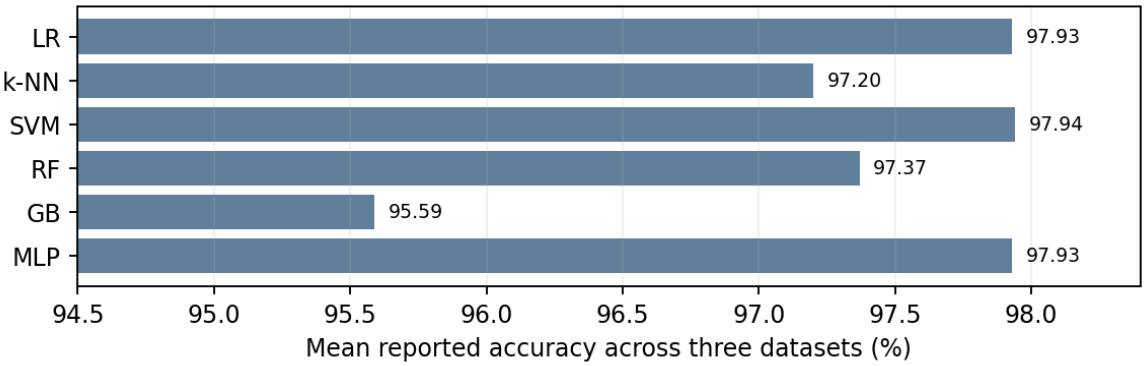
Selected source-reported corrected comparisons

Dataset	Pairwise comparison	Gap (pp)	t (df=29)	p
Digits	SVM vs gradient boosting	1.74	3.73	0.001
Digits	SVM vs logistic regression	1.21	2.93	0.007
Wine	Logistic regression vs gradient boosting	4.66	2.24	0.033
Digits	SVM vs MLP	0.70	1.96	0.060

The manuscript reports three comparisons with unadjusted $p < .05$ among 15 tests; only SVM versus gradient boosting on Digits is below the stated Bonferroni cut-off of .0033. No Breast Cancer comparison was reported significant (minimum $p = .123$). This narrow result does not establish equivalence of other algorithms, and selecting the best model before testing can complicate inferential interpretation.

4.1. Training cost. For Digits, the reported fit times were 0.003 s (k-NN), 0.051 s (SVM), 0.337 s (logistic regression), 0.704 s (MLP), 0.767 s (random forest), and 6.726 s (gradient boosting). Dividing 6.726 by 0.051 yields approximately 132 for this recorded benchmark; the ratio is not an inferential estimate of hardware-independent speed. In particular, k-NN defers much computational work to prediction, so low fit time alone cannot establish the lowest total deployment cost.

DOI: 10.65322/j.srd.05192026.68



1-rasm. Equal-weight mean accuracy calculated from the published dataset-level scores

4.2. Sample-size sensitivity

On Digits, the source learning curves report validation accuracy of 75.1% (logistic regression), 72.8% (SVM), and 72.6% (random forest) at 71 training cases, increasing to 97.1%, 98.4%, and 97.9%, respectively, at 1,437 cases. These curves suggest a sample-size-dependent ranking within this specific setting, not a universal threshold at which nonlinear models become preferable.

Discussion. The three datasets show no single algorithm winning all tasks, consistent with the need for empirical rather than a priori model selection. However, the comparison is neither a comprehensive algorithm survey nor evidence that nonlinear models or tree ensembles are generally inferior. Differences in sample size, signal geometry, data preprocessing and hyperparameter configuration may change relative results [4].

The reported gradient boosting settings are a particular sklearn implementation and should not be generalised to tuned boosting frameworks. The observation that logistic regression performs well on Breast Cancer and Wine may be compatible with relatively accessible class boundaries, but linear separability was not formally measured. The finding that a higher-capacity SVM performs better on larger Digits subsamples likewise remains an interpretation of learning curves, not an experimental identification of a causal bias–variance mechanism.

5.1. Statistical and application-level implications

The analysis should separate descriptive differences from formal conclusions. Although corrected pairwise testing is more conservative than an ordinary paired t-test on overlapping splits, p-values remain contingent on assumptions, code execution details and the chosen family of comparisons. The Friedman test with Nemenyi post-hoc comparisons can be appropriate when sufficient independent datasets form the sampling unit; with only three benchmarks, a critical-difference diagram would risk unjustified certainty [7].

For the Breast Cancer dataset, accuracy and macro-F1 alone cannot describe asymmetric diagnostic errors. Class-specific sensitivity, specificity, confusion matrices, and potentially calibrated performance would be needed to evaluate decision consequences. They cannot be reconstructed from the published means, so no medical sensitivity or specificity figures are reported. The educational dataset should not be presented as validation of a clinical diagnostic device.

5.2. Limitations and route to a stronger study

DOI: 10.65322/j.srd.05192026.68

The main limitation is experimental rather than editorial: three familiar teaching datasets, default or near-default hyperparameters, one reported computational environment, and no deposited executable code or raw folds. There is also no external validation cohort, prospective evaluation, calibration analysis, performance fairness audit or uncertainty interval for timing. These restrictions substantially constrain original scientific contribution.

A follow-up study should pre-register its model-selection criterion, compare tuned algorithms under nested CV with all preprocessing inside pipelines, enlarge the collection of genuinely independent datasets, include a journal-relevant applied dataset when permitted, and report paired effects with confidence intervals and multiplicity control. End-to-end training plus prediction latency, memory and repeat-timing dispersion should be measured on documented hardware. These are recommendations for future work and are not claimed as completed here.

Conclusion. RQ1: The best descriptive classifier varied by dataset–logistic regression led on Breast Cancer and Wine, whereas RBF-SVM led on Digits. RQ2: Of 15 reported corrected pairwise tests, three had unadjusted $p < .05$ and only one remained below the stated Bonferroni threshold; these claims await replication from raw fold results. RQ3: Training costs varied substantially on the recorded machine, while Digits learning curves exhibited changing relative model performance as the available training set increased.

Accordingly, repeated stratified validation can provide a useful starting benchmark when coupled with clear preprocessing, uncertainty reporting and documented computation. The present observations warrant baseline recommendations–not claims of algorithmic novelty, clinical readiness, statistical equivalence, or general superiority.

Data availability, code availability and declarations

The datasets are available through scikit-learn and associated public repositories. The numerical results in this revision derive from the supplied original article; per-fold scores, original executable code, verified repository, environment manifest and hardware details were not provided. Independent regeneration of the reported values has therefore not been established. Any new code deposit and author contact information must be confirmed by the authors before submission.

Funding, conflicts of interest, individual author contributions, ethics declarations and corresponding-author telephone details were not stated in the source and must be completed by the authors in accordance with the target journal's policy.

References

1. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.

2. Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the 14th International Joint Conference on Artificial Intelligence (Vol. 2, pp. 1137–1143). Morgan Kaufmann.

3. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, A., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

DOI: 10.65322/j.srd.05192026.68

4. Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? Advances in Neural Information Processing Systems, 35, 507–520.

5. Cawley, G. C., & Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. Journal of Machine Learning Research, 11, 2079–2107. https://jmlr.org/papers/v11/cawley10a.html

6. Nadeau, C., & Bengio, Y. (2003). Inference for the generalization error. Machine Learning, 52, 239–281. https://doi.org/10.1023/A:1024068626366

7. Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research, 7, 1–30. https://jmlr.org/papers/v7/demsar06a.html