JOURNAL OF
Emerging Paradigms
in Computing Systems
JEPCS

Journal of Emerging Paradigms in Computing Systems (JEPCS) | Volume 02, Issue No. 01, 2026

ARTICLE TYPE: Research Article

Article ID: JEPCS-2026-0201-0061

How Duplicate Samples Inflate Machine-Learning Evaluation: A Controlled Cross-Dataset Study of Train-Test Leakage

Ehsan Sarwar[1,*], Ayesha Ali[1]

[1] Departement of Computer Science, The University of Lahore, Lahore, Pakistan.

*Corresponding author: mehsan.sarwaruol@gmail.com

Contributing authors: mehsan.sarwaruol@gmail.com, ayeshaali453.uol@gmail.com

Received: October 11, 2025 **Accepted:** December 02, 2025

Abstract- Machine-learning evaluation assumes that the test set contains observations that were not available during model fitting. Exact duplicates that cross the train-test boundary break that assumption and can raise reported performance without improving generalization to new observations. This study measures that effect under controlled conditions. Four public classification datasets and one synthetic dataset are evaluated with logistic regression, 5-nearest neighbors, a decision tree, and a random forest. For each dataset, ten classified 70/30 train-test splits are created. The training partition is left unchanged, while 0, 5, 10, 20, 30, or 50 percent of the test records are replaced by exact class-matched copies of training records. This design isolates evaluation contamination from changes in training data. At 50 percent overlap, the mean inflation is 1.60 percentage points for logistic regression and 1.70 for 5-nearest neighbors. It rises to 7.15 points for the decision tree and 3.49 for the random forest. The largest single condition occurs on the synthetic dataset, where decision-tree accuracy rises by 18.37 points. At the same leakage level, decision-tree and random-forest accuracy on the duplicated subset is 100 percent, while their accuracy on the remaining uncontaminated records is 84.94 and 92.97 percent, respectively. Across the 20 dataset-model combinations, clean-test error is strongly associated with inflation at 50 percent overlap (Spearman $\rho = 0.81$, $p = 1.67 \times 10^{-5}$). The results show that duplicate leakage does not affect every classifier equally and can be most misleading when a flexible model has modest clean-test performance.

Keywords- Data leakage; duplicate samples; machine learning; reproducibility; cross-validation.

1. Introduction

A classifier is expected to be evaluated on observations that were not used to fit it. This requirement is simple to state, yet it is easy to violate when a dataset contains repeated records, repeated entities, near-duplicates, or preprocessing steps that use information from the test partition. Once train and test data are no longer independent, the reported score no longer measures the intended generalization problem. It measures performance on a mixture of new observations and information that has already entered the learning process.

Leakage has been discussed in data mining for more than a decade. Kaufman et al. [1] formalized leakage as information that would not legitimately be available at prediction time. More recent work has shown that the problem is not confined to competitions or poorly documented processes. Kapoor and Narayanan [2] reviewed leakage across scientific applications and found that it contributed to overoptimistic results in many fields. Rosenblatt et al. [3] reached a related conclusion in connectome-based prediction: repeated subjects and leaky feature selection could materially raise measured performance. These studies differ in domain, but they share the same evaluation concern. The test data must remain genuinely external to model development.

Exact duplicate leakage is a narrow but useful case to study since the contamination mechanism can be controlled exactly . A flexible classifier, when faced with a training-set record reappearing in the test set, may simply echo the training decision rather than generalizing from the surrounding structure. A decision tree can classify the duplicate with the same fitted partitions; an ensemble can inherit the same benefit across many trees. However , a linear model can not memorize an arbitrary sample like that . The effect of duplication should therefore depend on both the overlap and the family of models.

The literature contains strong warnings about leakage. The practical measurement question is: how much can exact overlap inflate common classification metrics, and does the answer differ across models and datasets? A controlled experiment can answer that question without mixing a number of leakage mechanisms. The training set can be fixed, the test set can be contaminated only and one can specify directly the overlap fraction. This design isolates the effect of duplicate evaluation samples from the effects of resampling, feature selection, hyperparameter tuning, or changes in the training distribution.

This design is used in this study. We evaluate four common classifiers on four public classification datasets and one synthetic dataset. Each dataset is trained with ten classified splits. For each split, the fitted model is kept constant and the test overlap is increased from 0 to 50 percent. We report accuracy, macro-F1, balanced accuracy and Matthews correlation coefficient (MCC). The analysis further distinguishes predictions on duplicated test records from predictions on the rest of uncontaminated records.

The experiment is guided by three questions:

- RQ1: How does increasing the fraction of train-test overlap impact the reported classification performance?
- RQ2: Are memory-capable tree and neighbor methods more sensitive to duplicate leakage than a linear classifier?
- RQ3: When clean-test performance is lower, does leakage inflation increase?

This paper does not model all types of leakage. It isolates one mechanism such that its effect is directly measured. The evidence we obtain offers a pragmatic warning for the construction of benchmarks, classroom experiments, applied machine-learning papers, and any workflow in which duplicate or repeated observations are able to cross the evaluation boundary.

2. Related Work

2.1. Leakage and Evaluation Integrity

The earliest broad treatments of leakage focused on information that enters the modeling process even though it would not be available at the time of prediction. Kaufman et al. [1] described leakage as a data-management and problem-formulation failure rather than only a coding error. Their learn-predict separation remains relevant to modern processes because many transformations can accidentally mix training and evaluation information.

Kapoor and Narayanan [2] extended the discussion to reproducibility in machine-learning-based science. Their review connects leakage to overly optimistic findings across multiple research domains and calls for documenting evaluation design in enough detail to reveal where information crosses the train-test boundary. The same principle is scikit-learn's recommended practice: split data before fitting transformations, and learn preprocessing parameters from training data onlycitepedregosa2011.

Rosenblatt et al. [3] provide direct empirical evidence of repeated-observation leakage. Their neuroimaging experiments explore several leakage mechanisms, including duplicated subjects, and find that repeated-subject leakage can significantly inflate measured predictive performance. The present study is more restricted. It does not combine domain-specific repeated measurements with a predictive process but rather introduces exact overlap at known rates and compares the response of four families of classifiers.

The same worry has reemerged in the evaluation of large models. Singh et al. [4] consider benchmark contamination in language-model testing, and Kandpal et al. [5] show that duplicated training sequences are strongly correlated with memorisation in language models. These are different settings than tabular classification, but they reinforce a common point: repeated examples can change what a benchmark score means.

2.2. Model Selection and Validation Bias

Leakage is one cause for optimism, but it is one part of a larger literature on validation bias. Varma and Simon [6] showed that using cross-validation for both model selection and error estimation can lead to biased error estimates unless an outer validation procedure is used. Cawley and Talbot [7] also studied overfitting to the model-selection criterion and the related selection bias. Dwork et al. [8] studied the repeated use of holdout data and proposed a mechanism to guarantee validity for adaptive analysis.

Other work addresses the uncertainty in the estimates of performance. Bengio and Grandvalet [9] have shown that there is no unbiased estimator of the variance of k-fold cross-validation that is valid for all cases. Nadeau and Bengio [10] examined inference for the generalisation error and the bias introduced by treating the cross-validation results

as independent. Even when the cross-validation procedure is correctly implemented, Varoquaux [11] showed that small samples can lead to a large uncertainty in the cross-validation estimates. These studies are on variance and selection, not duplicate leakage However, they do support the same methodological stance: the data split and the uncertainty structure have to be consistent with the claim made from the evaluation score.

Kohavi [12] compared cross-validation and bootstrap procedures for error estimation and Dietterich [13] studied statistical tests for comparing classifiers. Demsar [14] later reviewed nonparametric methods for comparing classifiers across multiple data sets. No claim is made in the current experiment that one classifier is better in general. We employ the classifiers as probes since their responses to exact overlap provide insight into the interaction of model structure and leakage.

2.3. Duplicate Data and Memorisation

Duplicate records impact learning in more than one way. The effects of repeated observations on class weighting, memorisation, privacy exposure, and frequency of gradients in training data can be changed. For instance, Kandpal et al. [5] demonstrate that duplication in language-model training data enhances memorisation, while deduplication decreases privacy risk. In the evaluation data this is another problem. The concern is not that a duplicate record will alter the fitted model, but that a test sample that should be unseen is already in the training set.

This distinction motivates the experiment employed here. No duplication or resampling is done on the training data. Exact duplicates are introduced into the test set only after the classifier has been trained. Thus any increase in the reported score is due to contamination of the evaluation set, not a changed model. The design also allows comparison of the accuracy on the overlapping subset with the accuracy on the rest of the records.

2.4. Classifiers and Metrics Used in the Study

The experiment contains four well-known classifiers with varied capacities. For a linear baseline we use logistic regression. The 5-nearest-neighbor classifier is an example of instance-based learning, where the nearest training examples are used for prediction according to the nearest-neighbor principle introduced by Cover and Hart [15]. A single decision tree is a partition model with high capacity. Random forest is the average of many randomised trees, and usually gives a smooth generalisation than a single tree [16].

As leakage inflation is easiest to interpret in percentage points, accuracy is maintained. Macro-F1 and balanced accuracy are added so that multiclass and uneven class distributions do not get lost behind a single majority weighted score. MCC is also reported. Chicco and Jurman [17] describe its usefulness as a summary measure for classification, in particular when the class frequencies are unequal.

3. Materials and Methods

3.1. Datasets

We use five datasets. Four are public benchmarks, available through scikit-learn [18] and the fifth is generated with `make_classification`, so that the experiment also contains a lower-separability problem. The experiment uses the dimensions reported in Table 1.

The Breast Cancer Wisconsin Diagnostic dataset consists of 569 samples and 30 real-valued features [19]. The wine dataset has 178 samples, 13 chemical measurements and 3 classes [20]. The Digits data in scikit-learn is derived from the optical handwritten-digit benchmark [21] and has 1,797 samples with 64 pixel-count features. Iris has 150 observations, four measurements and three species. The dataset dates back to Fisher's discriminant-analysis study [22]. The synthetic data set has 2,500 samples, 30 features, three classes, moderate class overlap and 4 percent label noise. It is on the list because a near ceiling clean benchmark has little room for visible score inflation.

Table 1: *Datasets used in the controlled leakage experiment.*

Dataset	Samples	Features	Classes	Smallest class
Breast Cancer	569	30	2	212
Wine	178	13	3	48
Digits	1,797	64	10	174
Iris	150	4	3	50
Synthetic-Overlap	2,500	30	3	523

3.2. Train-Test Split

Each dataset was classified, with a 70/30 split. We use 10 random seeds, from 0 to 9. Classification maintains the class proportions of each split. The training partition is given to the classifier and is not modified any further. This

fixed-training design matters because the experiment is designed to isolate evaluation leakage, not the effect of duplicate training records.

For logistic regression and 5-NN, the numerical features are standardised in a scikit-learn process. The scaler is fitted on the training partition and then applied on the test partition. The unscaled features are fed into the decision tree and random forest. The parameters are shown in Table 2.

Table 2: *Classifier settings.*

Classifier	Configuration
Logistic regression	StandardScaler followed by logistic regression; LBFGS solver; maximum 2,500 iterations.
5-nearest neighbors	StandardScaler followed by k-NN with $k = 5$.
Decision tree	Standard decision-tree classifier with split-specific random seed.
Random forest	60 trees; square-root feature sampling; split-specific random seed; all cores enabled.

3.3. Controlled Duplicate Leakage

The four classifiers are fit and the clean test set is copied six times. Then train-test overlap is introduced at 0, 5, 10, 20, 30 and 50 percent of the training data. The requested number of contaminated test positions is distributed proportionally across the classes. For each selected test row, we replace it with the corresponding feature vector sampled from the training rows of the same class. Its label is still class consistent.

The procedure keeps the test-set size and roughly the class balance but controls the fraction of records with feature vectors that have been seen before in training. It does not alter the model parameters. The sequence is summarised in Figure 1.

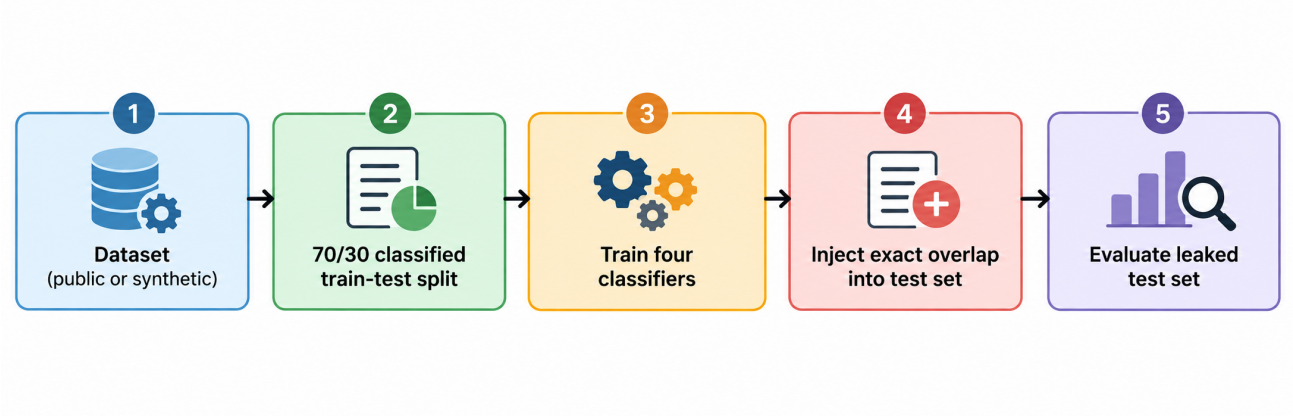


Figure 1: *Controlled train-test contamination procedure. The fitted classifier is fixed within a split; only the test representation is changed across overlap levels.*

Let A_0 denote accuracy on the clean test partition and A_λ accuracy when a fraction λ of test records are exact copies of training records. Accuracy inflation is defined as

$$\Delta A(\lambda) = A_\lambda - A_0. \tag{1}$$

A positive value means that the contaminated evaluation reports a higher score than the clean evaluation of the same fitted model.

3.4. Metrics and Secondary Analysis

The primary result is an increase in accuracy, because it can be directly interpreted in percentage points. Balanced Accuracy, MCC and Macro-F1 are also displayed. For each contaminated test set the predictions are split into two sets: exact-overlap records and records that are still uncontaminated. This division lets us see if the high total score is simply because of the repeated examples.

The results are averaged over the ten random splits. To investigate RQ3, we compare clean-test error $(1 - A_0)$ with accuracy inflation at 50 percent overlap for the 20 dataset-model combinations. Since neither quantity is assumed to follow a linear normal model, we use the Spearman rank correlation . In heterogeneous datasets the analysis is descriptive; the correlation is not interpreted as law.

4. Results

4.1. Clean Baselines

The clean baseline scores vary widely across datasets and models. Table 3 shows the mean accuracy over ten folds. Most models have high baseline scores for Wine, Breast Cancer and Digits. The synthetic dataset is designed to be harder. The average accuracy varies from 0.6215 for logistic regression to 0.7848 for random forest. The decision-tree has also a poorer performance than the corresponding ensemble result on Digits and Synthetic-Overlap.

Table 3: *Mean clean-test accuracy over ten classified splits.*

Dataset	LR	5-NN	Tree	RF
Breast Cancer	0.9719	0.9632	0.9322	0.9591
Wine	0.9815	0.9500	0.9167	0.9870
Digits	0.9696	0.9731	0.8415	0.9720
Iris	0.9489	0.9356	0.9467	0.9467
Synthetic-Overlap	0.6215	0.7504	0.6284	0.7848

The interpretation of the leakage experiment requires these baselines. Since the clean accuracy is already close to one there is not much numerical room for a contaminated test set to increase the score. A weaker baseline leaves more errors that can be replaced by examples the model has seen.

4.2. Inflation as Exact Overlap Increases

The average accuracy inflation over the five datasets and ten splits is shown in Fig. 2. The four model families do not react the same. For a 50% overlap, mean inflation is 1.60 percentage points for logistic regression, 1.70 for 5-NN, 7.15 for the decision tree and 3.49 for the random forest. For the two tree-based models, the increase is monotonic on the whole.

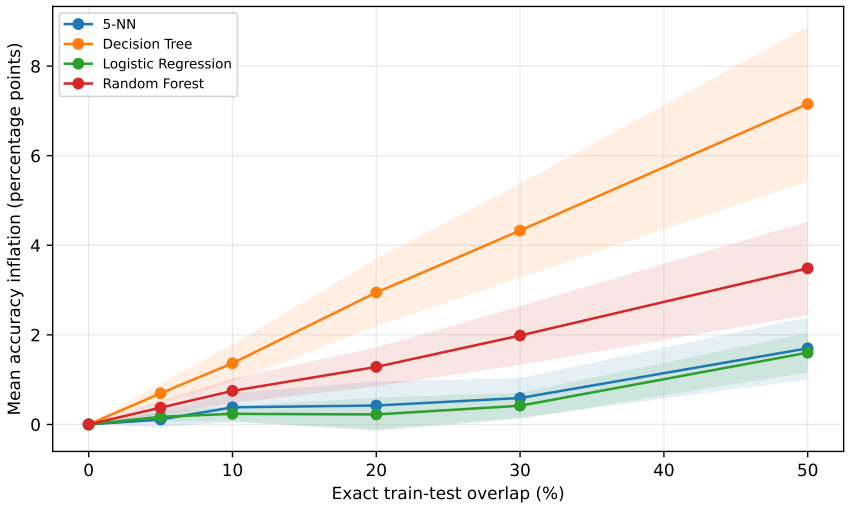


Figure 2: *Mean accuracy inflation as exact train-test overlap increases. Shaded regions show mean* ±1.96 *standard errors over the 50 pooled dataset-split observations for each model and overlap level.*

Table 4 gives the mean inflation at each contamination level. At only 10 percent overlap, the decision tree gains 1.37 points on average and random forest gains 0.75 points. The corresponding changes for logistic regression and 5-NN are smaller. At 50 percent overlap, the gap between the model families is much clearer.

Table 4: *Mean accuracy inflation in percentage points across five datasets and ten splits.*

Overlap	LR	5-NN	Tree	RF
5%	0.17	0.11	0.69	0.38
10%	0.24	0.38	1.37	0.75
20%	0.22	0.42	2.95	1.28
30%	0.42	0.59	4.33	1.99
50%	1.60	1.70	7.15	3.49

The small non-monotonic changes in LR and 5-NN at intermediate overlap levels are caused by which clean test records are replaced in a finite sample. Exact duplicates are usually easier for these models as well, but an injected record

is not guaranteed to correct an error. The tree-based pattern is more direct because exact training vectors are classified correctly by the fitted tree and forest in this experiment.

4.3. Dataset Dependence

The average results hide large differences between datasets. Figure 3 reports inflation for every dataset-model combination at 50 percent overlap. The largest effect occurs for the decision tree on Synthetic-Overlap: accuracy rises by 18.37 percentage points, from 0.6284 to 0.8121. Random forest on the same dataset gains 10.43 points. The decision tree on Digits gains 8.09 points.

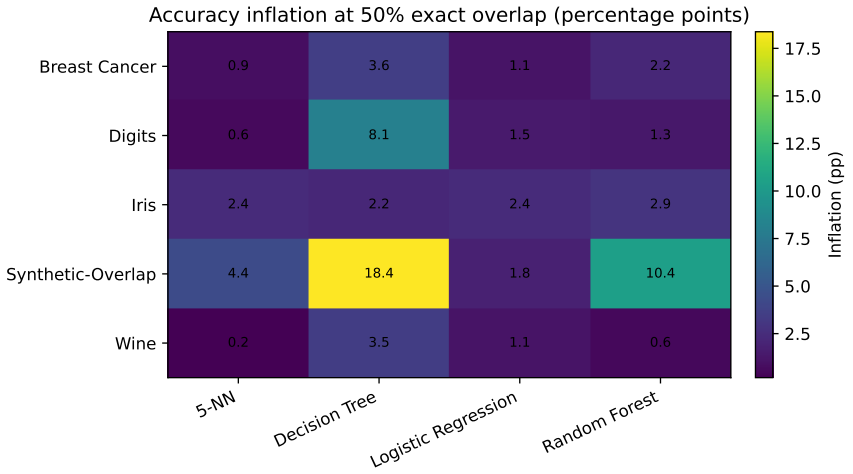


Figure 3: *Accuracy inflation at 50 percent exact overlap for each dataset-model combination. Values are percentage points averaged over ten splits.*

Several high-performing baselines move much less. Wine random forest gains 0.56 points and Wine 5-NN gains 0.19 points. 5-NN of Digits gains 0.63 points. But these small differences do not mean that overlap is safe. They have a ceiling effect: if the clean accuracy is already around 0.97 to 0.99, then a corrupted score cannot be improved by many points.

The synthetic data set makes the distortion easier to see. Figure 4 shows reported accuracy as overlap increases. The decision-tree curve ranges from 62.84 percent at no overlap to 81.21 percent at 50 percent overlap. Within a split, all six evaluations have the same fitted tree. Random forest jumps from 78.48 percent to 88.91 percent. The test composition changes and the score changes, not because the classifier improved.

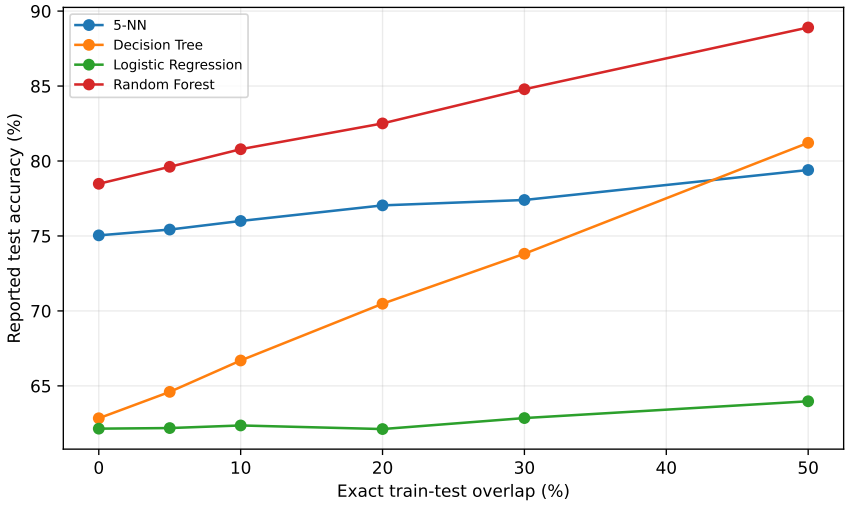


Figure 4: *Reported test accuracy on Synthetic-Overlap as exact train-test overlap increases. Models are not retrained between overlap levels within a split.*

4.4. Performance on Overlapping and Uncontaminated Records

The subset analysis explains the aggregate inflation. At 50 percent overlap, the decision tree and random forest classify the exact-overlap subset with 100 percent mean accuracy. Their mean accuracy on the records that remain uncontaminated is 84.94 and 92.97 percent, respectively. The total test score combines these two different evaluation conditions.

For 5-NN, mean accuracy is 95.24 percent on the overlapping subset and 91.06 percent on the uncontaminated subset. Logistic regression shows a smaller separation: 92.52 versus 90.43 percent. Figure 5 summarizes the comparison.

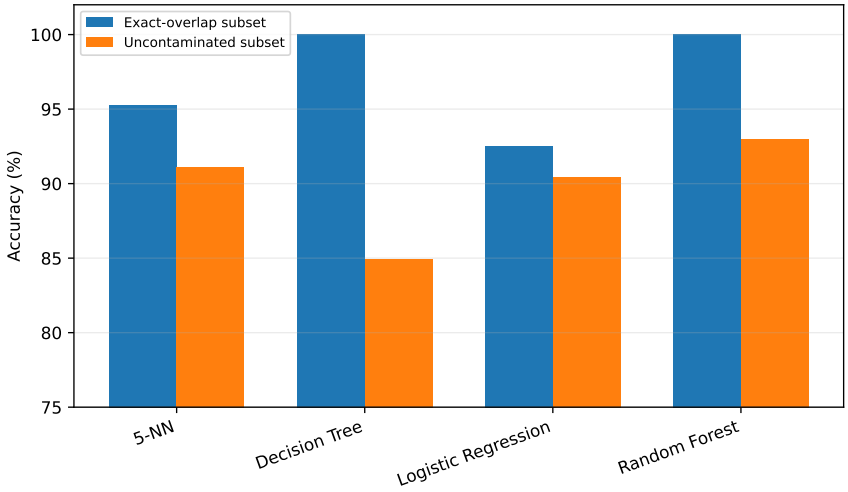


Figure 5: *Mean accuracy on exact-overlap and uncontaminated test records at 50 percent train-test overlap.*

This result addresses RQ2. The largest leakage response occurs for the models that can fit highly local or partition-specific structure. The decision tree reproduces every injected training example correctly, and the random forest does the same in the observed runs. The linear model gains much less because an exact repeated row does not create a separate memorized rule for that sample.

4.5. Baseline Difficulty and Leakage Inflation

Across the 20 dataset-model means, clean-test error is positively associated with inflation at 50 percent overlap. Spearman correlation is $\rho = 0.81$ with $p = 1.67 \times 10^{-5}$. Figure 6 shows the relationship. The strongest points are the synthetic decision tree, synthetic random forest, and Digits decision tree, all of which have larger clean errors and larger leakage gains.

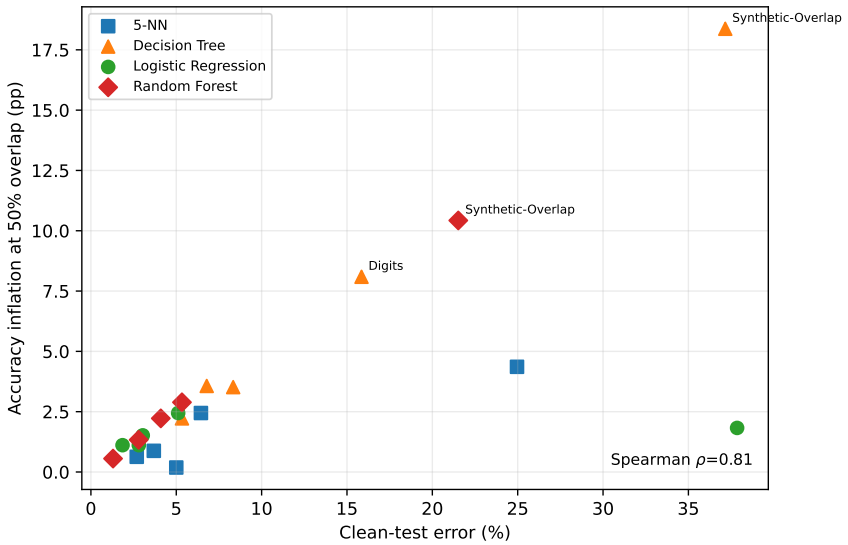


Figure 6: *Clean-test error versus accuracy inflation at 50 percent overlap across the 20 dataset-model combinations.*

The correlation should be interpreted within this experiment. It is partly a mathematical opportunity effect: a model that already classifies nearly every clean sample correctly has fewer errors that duplicated training samples can replace. It is also model capacity dependent. The result thus supports RQ3 for the evaluated conditions but does not show that baseline error alone drives leakage sensitivity.

5. Discussion

5.1. What The Contaminated Score Measures

A contaminated test score does not provide a noisy estimate of clean generalisation. It accounts for a different mix of cases. Some records are really unseen, some other records are exact feature vectors available at training time. The aggregate score depends directly on the mixture proportion of these groups when they are of different accuracies.

This is most apparent for the tree models. With 50 percent overlap half of the test set is replaced by training copies of the same class. In the runs that we have seen, the decision tree and the random forest classify that half perfectly. Thus their overall accuracy increases even if accuracy on the remaining unseen cases remains close to its starting level. A reader who only sees the aggregated test score cannot infer how much of the reported performance is due to true generalisation.

5.2. Why is model family important

The four classifiers are presented with the same contaminated records, but they behave differently. A decision tree can reproduce fitted partition decisions on an exact training vector. Random forests average many trees, but the duplicated sample still has the benefit of being in the training data. The instance-based 5-NN is also sensitive to local repetitions, although one exact neighbour does not guarantee the majority vote of five neighbours. Logistic regression fits a global linear boundary. A repeated test vector from training does not have to predict correctly.

The difference provides useful input for the design of the audit. A little increase in one model does not prove the dataset is overlap free. A linear model at the ceiling may show little inflation even if contamination is present, and a benchmark at the ceiling may have little headroom. The overlap is to be checked directly, not inferred from unusually high accuracy.

5.3. Practical assessment procedure

A defensible workflow begins before you fit the model. First determine the unit of independence. If all related observations have to stay in one partition, multiple rows can be used for one patient, device, person, file or event. When a stable entity identifier is available, group-aware splitting is better than random row-wise splitting. The problem is illustrated by Rosenblatt et al. [3] with repeated subjects, but extends to other repeated-measurement datasets.

Exact match can be audited by hashing canonicalized rows or stable record identifiers. Finding near-duplicates requires domain-specific similarity measures and should not be confused with exact hashing. Preprocessing, feature selection, imputation and scaling should then be fitted only on the training data, in line with leakage-avoidance guidancecitekaufman2012,pedregosa2011. If overlap is found after evaluation, the contaminated and uncontaminated results should be reported separately, rather than silently deleting inconvenient samples after model selection.

The experiment provides a simple checklist for reporting:

- define the independent entity represented by a row
- split before any learned preprocessing or resampling
- audit exact train-test overlap and document the method
- use group-aware partitions when repeated entities exist
- report number or proportion of overlapping records if they cannot be removed
- avoid interpreting a high contaminated score as evidence of deployment generalisation.

5.4. Relevance to reproducibility

Leakage can affect more than one experiment. A benchmark that is tainted can also have an impact on later work. One can take its score as a baseline, as a tuning target, or as evidence that the prediction problem is nearly solved. Kapoor and Narayanan [2] discuss this broader reproducibility concern across scientific machine learning. The present experiment gives a controlled numerical picture of one common mechanism, exact duplicate overlap.

The result also highlights why duplicate checking is not just a data-cleaning step. Synthetic-Overlap is the same decision tree goes from 62.84 to 81.21 percent reported accuracy as overlap reaches 50 percent. The learned classifier does not improve anything. The difference is entirely due to the evaluation data. This distinction matters when comparing benchmark results across papers.

6. Limitations

The experiment tells apart identical copies. Real datasets may contain near-duplicates, repeated entities with different measurements, temporal leakage, variables derived from the target, or preprocessing leakage. These mechanisms may act differently and are not quantified here.

Four out of the five datasets are small to medium-sized teaching and benchmark datasets. Their value is diversity of class structure and reproducibility, not realism for every modern application. The synthetic data set is a harder classification problem, but it is generated from a controlled statistical model. The results should thus be interpreted as evidence about the leakage mechanism, and not as a ranking of classifiers for production tasks.

Only four classifiers are tested. They were chosen to represent linear, instance-based, single-tree and ensemble behaviour. Different neural networks, boosting methods, support-vector machines or specialised deep architectures may respond differently. The model dependent sensitivity is one of the results, so the main result does not require the sensitivity of all classifiers to be the same.

The contamination procedure substitutes training copies of the same class for test observations. This preserves class balance and gives direct control over the overlap fraction, but it is an experimental intervention rather than a reconstruction of how duplicates enter into a real data process. In practice, duplicates may arise before splitting, may have different frequencies, or may be connected by an entity identifier rather than by exact feature equality.

Finally, the ten random splits within a dataset are not treated as ten independent scientific datasets. They are used to reduce split-specific noise. The cross-dataset Spearman analysis uses the 20 dataset-model means and is reported as a descriptive relationship for the evaluated conditions.

7. Conclusion

Exact train-test duplicates can raise machine-learning evaluation scores without improving the fitted model. The size of that distortion depends on the classifier and on the underlying clean-test difficulty. At 50 percent exact overlap, mean accuracy rises by 7.15 percentage points for the decision tree and 3.49 points for random forest. Logistic regression and 5-NN rise by 1.60 and 1.70 points, respectively. The most affected condition is the decision tree on the synthetic dataset, where the reported score increases by 18.37 points.

The subset analysis explains why. Decision-tree and random-forest accuracy on duplicated test records reaches 100 percent at the 50 percent overlap condition, while accuracy on the remaining uncontaminated records is lower. The resulting aggregate score is a mixture of memorised and truly unseen observations. The relationship between inflation and clean-test error is also robust across dataset-model combinations ($\rho = 0.81$), which suggests that leakage can be especially misleading when baseline performance leaves room for apparent improvement.

The practical implication is straightforward: one must check evaluation integrity at the data-split level, not infer from the final metric. We propose cheap precautions such as exact-overlap audits, entity-aware grouping, train-only preprocessing, and separate reporting of contaminated and uncontaminated performance. They do not address all types of leakage, but they stop a basic data-management error from being mistaken for model progress.

Author Declarations

Author Contributions (CRediT): Ehsan Sarwar: Conceptualization, Methodology, Software, Investigation; Ayesha Ali: Data Curation, Formal Analysis, Validation, Visualization, Writing - Original Draft, Writing - Review & Editing. The authors have read and approved the submitted version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The public benchmark datasets used in this study are distributed with scikit-learn and originate from the UCI Machine Learning Repository.

Acknowledgments: No acknowledgments are reported.

Conflicts of Interest: The authors declares no conflict of interest.

Ethics Statement: Not applicable.

Declaration of Generative AI: An AI-assisted tool was used during manuscript preparation for code support, literature organization, and language editing. The authors reviewed the manuscript text and remains responsible for the accuracy and integrity of the article.

References

[1] S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman, "Leakage in data mining: Formulation, detection, and avoidance," *ACM Transactions on Knowledge Discovery from Data*, vol. 6, no. 4, pp. 15:1–15:21, 2012.

[2] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9,

p. 100804, 2023.
[3] J. D. Rosenblatt *et al.*, "Data leakage inflates prediction performance in connectome-based machine learning models," *Nature Communications*, vol. 15, 2024.
[4] A. K. Singh, M. Y. Kocyigit, A. Poulton, D. Esiobu, M. Lomeli, G. Szilvasy, and D. Hupkes, "Evaluation data contamination in llms: How do we measure it and when does it matter?" *arXiv preprint arXiv:2411.03923*, 2024.
[5] N. Kandpal, E. Wallace, and C. Raffel, "Deduplicating training data mitigates privacy risks in language models," in *Proceedings of the 39th International Conference on Machine Learning*, vol. 162, 2022, pp. 10 697–10 707.
[6] S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection," *BMC Bioinformatics*, vol. 7, p. 91, 2006.
[7] G. C. Cawley and N. L. C. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *Journal of Machine Learning Research*, vol. 11, pp. 2079–2107, 2010.
[8] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth, "The reusable holdout: Preserving validity in adaptive data analysis," *Science*, vol. 349, no. 6248, pp. 636–638, 2015.
[9] Y. Bengio and Y. Grandvalet, "No unbiased estimator of the variance of k-fold cross-validation," *Journal of Machine Learning Research*, vol. 5, pp. 1089–1105, 2004.
[10] C. Nadeau and Y. Bengio, "Inference for the generalization error," *Machine Learning*, vol. 52, pp. 239–281, 2003.
[11] G. Varoquaux, "Cross-validation failure: Small sample sizes lead to large error bars," *NeuroImage*, vol. 180, pp. 68–77, 2018.
[12] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1995, pp. 1137–1145.
[13] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Computation*, vol. 10, no. 7, pp. 1895–1923, 1998.
[14] J. Demsar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.
[15] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.
[16] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.
[17] D. Chicco and G. Jurman, "The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, p. 6, 2020.
[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. VanderPlas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
[19] W. Wolberg, O. Mangasarian, N. Street, and W. Street, "Breast cancer wisconsin (diagnostic)," UCI Machine Learning Repository, 1993.
[20] S. Aeberhard and M. Forina, "Wine," UCI Machine Learning Repository, 1992.
[21] E. Alpaydin and C. Kaynak, "Optical recognition of handwritten digits," UCI Machine Learning Repository, 1998.
[22] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.