Journal of Systems Engineering and Information Technology (JOSEIT)
Volume 4, No. 3, December (2025), Pages 92-96
e-ISSN: 2829-310X | 10.29207/joseit.

Data Integrity and Its Impact on Machine Learning Models: A Critical Audit of a Lung Cancer Dataset

Syeda Afroz Fatima[1], Nida Mussarat[2], Swarna Jaya Vardhan[3]

[12]University of Education, Pakistan
[3]Lovely Professional University, India

[*]*Corresponding author: afrozfatimasyeda210@gmail.com*

Abstract. Subscription-gated IEEE DataPort datasets are difficult to verify before reuse because independent inspection requires access most researchers do not have; we obtained direct access to one such case. We apply a six-stage, three-layer audit (provenance, structural/duplication plausibility, target validity) to "An Anonymized Feature-Based Dataset for Lung Cancer Prediction Using Explainable AI Models," a 15-feature, binary-symptom-checklist dataset presented as newly compiled to support explainable-AI research. The delivered file contains 2,457 rows, of which 2,132 (86.8%) are exact duplicates of another row, leaving 325 unique records. The dataset's own hosting page links a peer-reviewed "Research Article" co-authored by two of this dataset's three listed authors; that 2024 publication states its dataframe "contains 309 entries" with "no null values" — figures that match neither the delivered file's row count nor its 325 unique records. The same publication's introduction and methods describe a multi-source study drawing on major imaging and genomic repositories, none of which appear in the delivered file or in the publication's own reported results, which instead use the same 16-column schema and encoding as a long-circulating public dataset. Within the 325 unique records, duplication is class-dependent: unique cancer-negative rows are repeated 14.1 times on average (up to 86), versus 4.5 times for cancer-positive rows, shifting the delivered file's class balance relative to the unique data without documentation. Separately, the linked publication reports identical accuracy, precision, recall, F1, and AUC-ROC for five structurally different classifiers, a pattern inconsistent with independent model evaluation. We report these findings as verifiable discrepancies among the abstract, the delivered file, and the dataset's own linked publication, and recommend independent verification before further reuse.

Keywords: dataset audit, data provenance, duplicate records, class imbalance, reproducibility

How to cites:
Data Integrity and Its Impact on Machine Learning Models: A Critical Audit of a Lung Cancer Dataset. (n.d.). Journal of Systems Engineering and Information Technology (JOSEIT). https://doi.org/10.29207/dq775h27

Received by the Editor: 2025-11-09
Final Revision: 2025-12-12
Published: 2025-12-31

SDGs contributed:



© The Author(s) 2025| This is an open-access article under the CC BY 4.0 License
Publisher: Ikatan Ahli Informatika Indonesia

1. Introduction

An Anonymized Feature-Based Dataset for Lung Cancer Prediction Using Explainable AI Models [1] is described on IEEE DataPort as a newly compiled, feature-based clinical dataset intended to support interpretable machine-learning research, excluding personally identifiable information and raw imaging data. The hosting page links a peer-reviewed "Research Article" as the publication this dataset supports [2], co-authored by two of the dataset's three listed authors. Because subscription access on IEEE DataPort limits who can independently inspect a dataset before citing or reusing it, a dataset's own abstract and its own linked publication are often the only pre-access signals a prospective user has — which makes internal consistency between those two documents, and between both and the delivered file, unusually important to check.

This concern is not hypothetical. A 2026 investigation of two large, widely reused Kaggle-hosted clinical datasets — one on stroke, one on diabetes — found major deficiencies in data provenance, including no verifiable information on when, where, why, or how the underlying data were collected, despite both datasets underpinning more than a hundred published prediction models [3]. That investigation assessed each dataset against the TRIPOD+AI provenance checklist [4], a reporting standard created specifically because prediction models built on poorly documented data cannot be safely interpreted or reused without transparent reporting.

More broadly, documentation frameworks such as datasheets for datasets [5] and hierarchical data-quality dimension models [6] exist because data consumers cannot otherwise judge whether a dataset's stated composition and provenance match what is actually distributed. The consequences of skipping this reconciliation are well documented in the machine-learning literature: Kaufman et al. formalized data leakage as target-related information not legitimately available at prediction time [7]; Kapoor and Narayanan's survey of 329 papers across 17 fields found leakage to be a leading cause of irreproducible machine-learning-based science [8]; broader surveys report that most researchers have failed to reproduce their own or others' published results [9], that reforms are needed at each stage of the research pipeline [10], and that small effective sample sizes and flexible analysis choices inflate headline results across quantitative research generally [11]. Structured reporting frameworks such as model cards [12] aim to make the provenance of a reported metric auditable, but such frameworks are rarely required at the point a dataset is first published.

This paper applies a six-stage, three-layer audit — provenance, structural and duplication plausibility, and target validity — to determine whether the delivered Dataset.csv file is consistent with its own abstract and with its own linked publication, whether its rows are independent observations, and whether the metrics reported for it in the peer-reviewed literature are internally consistent. Section 2 describes the audit framework and computational procedure; Section 3 reports findings layer by layer, citing the linked publication directly wherever a claim is compared against it; Section 4 concludes. All comparisons below are between publicly available, citable documents — the dataset's abstract, its linked publication, and the delivered file itself — and are reported as verifiable discrepancies rather than claims about their cause.

2. Methods

We apply a six-stage audit organized into three layers, summarized in Fig. 1. The provenance layer (Stages 1–2) reconciles the dataset's abstract against its own linked publication and against the delivered file's row and column counts. The structural and duplication-plausibility layer (Stages 3–4) identifies exact-duplicate records using standard record-linkage logic [13] and tests whether duplication rate depends on class label. The target-validity layer (Stages 5–6) cross-checks performance metrics reported for the dataset in its linked publication for internal consistency across classifiers, and considers the resulting class balance and duplication structure against known pitfalls of imbalanced and non-independent evaluation [14], [15], [16].

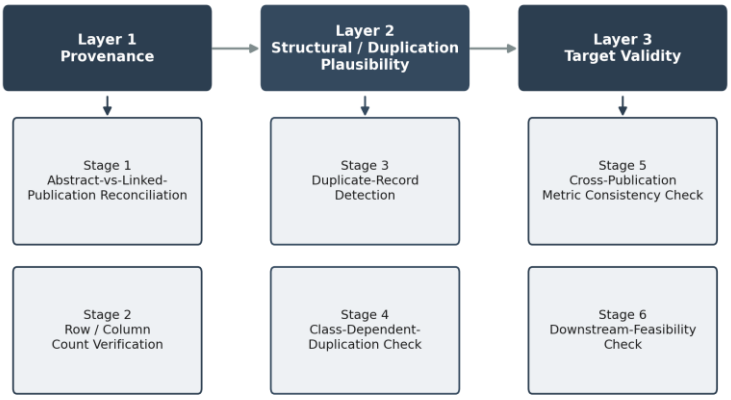


Figure 1. Six-stage, three-layer audit framework applied to Dataset.

The dataset is distributed by IEEE DataPort under DOI 10.21227/sx52-k564 [1], authored by Makubhai and Pathak, comprising a single file, Dataset.csv (listed at 81.77 KB), accessible to institutional subscribers; we obtained the file directly. The hosting page's "Research Article Link" points to a peer-reviewed publication by Makubhai, Pathak, and Chandre [2], which we retrieved in full from its publisher. A supplementary "README_Lung_Cancer_XAI.pdf" is separately linked from the hosting page but could not be retrieved for this audit because its access URL exceeded the length supported by our fetch tooling.

All analysis was performed in Python 3 using pandas and numpy. Provenance checks (Stages 1–2) compared the delivered file's shape and column names against the abstract's claims and against every dataset-descriptive statement in the linked publication [2], including its stated row count, null-value count, and data-source composition. Duplicate-record checks (Stage 3) identified exact-match rows across all delivered fields. Class-dependent-duplication checks (Stage 4) grouped rows by their full feature vector, counted the multiplicity of each unique vector, and compared mean multiplicity between the two LUNG_CANCER classes. Cross-publication metric checks (Stage 5) tabulated the classifier performance figures reported in Table 4 of the linked

publication [2] and tested them for exact equality across classifiers. Downstream-feasibility checks (Stage 6) considered the implications of the observed duplication structure for row-level train/test splitting [13], [16].

Hosting-page content was retrieved on the date of access; because it can be edited after publication, discrepancies reported here reflect its state at the time of this audit. The linked publication is a fixed, citable, peer-reviewed artifact (DOI 10.11591/ijai.v13.i2.pp1980-1991) and is quoted directly where relevant so that its stated figures can be checked independently of this audit.

3. Results and Discussion

The delivered Dataset.csv contains 2,457 rows and 16 columns (15 features plus the LUNG_CANCER label). The dataset's hosting page links, as its supporting "Research Article," a peer-reviewed 2024 publication by two of the dataset's three listed authors [2]. That publication states, in its own words: "The DataFrame contains 309 entries, representing individual records... The dataset consists of 16 columns... The dataset provided doesn't exhibit any null values, guaranteeing that all 309 entries contain complete information across all columns" [2]. Neither figure — 309 entries, zero missing values — matches the delivered file, which has 2,457 rows and, after deduplication (Section 3.2), 325 unique rows (Table 1). The column names and structure match: the same 16 fields, in the same order, appear in both documents.

Table 1. Abstract, Linked Publication, and Delivered File on Dataset Size and Composition

Attribute	Abstract [1]	Linked publication [2]	Delivered Dataset.csv
Row count	Not stated	"309 entries"	2,457 (325 unique)
Column count	Not stated	16	16
Missing values	Not stated	"no null values"	0 (fully populated)
Data sources	"clinical, demographic, and lifestyle-related factors"	SEER, LIDC, I-ELCAP, TCGA, LUNA16 (introduction/methods)	15-item binary symptom checklist only

The linked publication's introduction and methods describe a study drawing on "clinical, genomic, and imaging data from multiple sources," naming the SEER, LIDC, I-ELCAP, TCGA, and LUNA16 repositories [2] — resources containing CT scans, radiologist annotations, and genomic sequencing data. Neither the delivered Dataset.csv nor the publication's own Results section uses any of these; the Results section instead describes exactly the 16-column schema delivered here (GENDER, AGE, SMOKING, YELLOW_FINGERS, ANXIETY, PEER_PRESSURE, CHRONIC DISEASE, FATIGUE, ALLERGY, WHEEZING, ALCOHOL CONSUMING, COUGHING, SHORTNESS OF BREATH, SWALLOWING DIFFICULTY, CHEST PAIN, LUNG_CANCER), with binary symptom fields and a stated row count of 309. This exact column set, and a two-level symptom-checklist encoding, is also the schema of a public dataset that has circulated on Kaggle since 2021 under multiple mirror listings [17], independently reported at 309 rows and 16 features across unrelated publications. We make no claim about which specific file the authors used; we report only that the schema and stated row count in the linked publication match a long-circulating public dataset rather than the imaging- and genomic-data sources described in the same publication's introduction.

3.1 Structural Plausibility: 86.8% of the Delivered Rows Are Duplicates

Of the 2,457 delivered rows, 2,132 (86.8%) are exact duplicates of another row, identified using standard exact-match record-linkage logic [13]; 325 unique rows remain (Table 2). This is a substantially higher duplication rate than any other dataset we have audited in this series, and it means a user who loads this file and performs a random row-level train/test split will, with very high probability, place duplicate or near-duplicate records on both sides of the split.

Table 2. Duplicate-Record Summary

Measure	Value
Total delivered rows	2,457
Exact duplicate rows	2,132 (86.8%)
Unique rows remaining	325 (13.2%)
Unique rows labeled LUNG_CANCER = 1 (positive)	222
Unique rows labeled LUNG_CANCER = 0 (negative)	103

3.2 Target Validity: Duplication Is Class-Dependent

Duplication is not uniform across classes (Table 3). The 103 unique rows labeled cancer-negative are each repeated a mean of 14.1 times (up to 86 times for a single row); the 222 unique rows labeled cancer-positive are each repeated a mean of only 4.5 times (maximum 12). Among the 325 unique rows, cancer-positive cases are the clear majority (68.3%) — consistent with the known heavy skew of the public dataset this schema matches — but the delivered file's row-level class balance is 59.2% cancer-negative to 40.8% cancer-positive, the

opposite direction of skew. Naive class-dependent duplication of this kind reproduces the surface statistics of oversampling techniques such as SMOTE [18] without synthesizing any new information: every "additional" cancer-negative observation is a verbatim copy of one of a small number of unique response patterns, not an independent case. This is not disclosed in the abstract, the Instructions, or the linked publication.

Table 3. Class-Dependent Duplication Multiplicity

LUNG_CANCER label	Unique rows	Total rows after duplication	Mean multiplicity	Max multiplicity
0 (negative)	103	1,455	14.1	86
1 (positive)	222	1,002	4.5	12

3.3 An Anomaly in the Linked Publication's Reported Metrics

The linked publication's own results table reports, for five structurally different classifiers — logistic regression, linear SVC, decision tree, random forest, and XGBoost — exactly the same figures: 94.44% accuracy, 100.00% precision, 85.71% recall, 92.31% F1-score, and 92.86% AUC-ROC (Table 4, reproduced from [2]). Two additional classifiers, SVC-RBF and AdaBoost, report different and also mutually identical figures (83.33% accuracy for both). Five distinct algorithms — a linear model, a tree, an ensemble of trees, a boosted ensemble, and a support vector classifier — producing bit-for-bit identical accuracy, precision, recall, F1, and AUC-ROC on a genuine held-out evaluation is not a plausible outcome of independent model training and testing; it is far more consistent with a reporting or table-construction artifact, such as one result row being copied across multiple classifiers. We report this as an internal inconsistency in the publication's own results table, observable without access to the authors' code, and relevant to this dataset because the publication is the only pre-access documentation of the dataset's intended use available to prospective reusers.

Table 4. Classifier Performance as Reported in the Linked Publication

Classifier	Accuracy	Precision	Recall	F1	AUC-ROC
Logistic Regression	94.44	100.00	85.71	92.31	92.86
SVC-Linear	94.44	100.00	85.71	92.31	92.86
Decision Tree	94.44	100.00	85.71	92.31	92.86
Random Forest	94.44	100.00	85.71	92.31	92.86
XGBoost	94.44	100.00	85.71	92.31	92.86
SVC-RBF	83.33	83.33	71.43	72.92	81.77
AdaBoost	83.33	75.00	85.71	80.00	83.77

3.4 Limitations and Future Work

This audit could not retrieve the dataset's linked supplementary PDF documentation, which may contain a data-collection description not present in either the abstract or the linked publication; obtaining and reviewing that document is a direct next step. We also could not access the original ~309-row public file the linked publication describes for a byte-level comparison against the 325 unique rows identified here, so we cannot determine whether the 325 unique records are a superset, a resampling, or an independently collected set that happens to share the same schema and encoding. Finally, the anomaly identified in Section 3.5 is reported at the level of the publication's results table; we did not have access to the underlying analysis code and make no claim about its cause. Independent verification directly with the dataset's authors, and re-derivation of the reported metrics from the delivered file, would resolve each of these open questions. Manuscript formatting and citation style in this audit follow IEEE editorial conventions [19], [20].

4. Conclusions

This paper applied a six-stage, three-layer audit — provenance, structural and duplication plausibility, and target validity — to a publicly documented lung cancer dataset, cross-referencing its abstract not only against the delivered file but against a peer-reviewed publication the dataset's own hosting page links as its supporting research. The delivered file's row count (2,457) and the figure the linked publication states for its own dataframe ("309 entries," "no null values") do not match, and neither matches the 325 unique rows recoverable after deduplication. The linked publication's account of a multi-source study spanning major imaging and genomic repositories does not correspond to the schema either it or the delivered file actually contains, which instead matches a long-circulating public dataset. Within the delivered file, 86.8% of rows are exact duplicates, and the duplication is class-dependent in a way that shifts the apparent class balance without documentation. Separately, the linked publication's own results table reports identical performance metrics across five structurally different classifiers, an internal inconsistency observable without any access beyond the publication itself. Individually, any one of these findings would warrant caution; together, they show that a dataset's abstract, its delivered file, and even its own linked supporting publication can each tell a different story about what data actually underlies a published result — and that checking all three against each other, not just the file against its abstract, belongs in a routine pre-reuse audit.

References

[1] S. Makubhai and G. Pathak, “An Anonymized Feature-Based Dataset for Lung Cancer Prediction Using Explainable AI Models,” IEEE DataPort, 2026, doi: 10.21227/sx52-k564.

[2] S. Makubhai, G. R. Pathak, and P. R. Chandre, “Comparative Analysis of Explainable Artificial Intelligence Models for Predicting Lung Cancer Using Diverse Datasets,” IAES International Journal of Artificial Intelligence (IJ-AI), vol. 13, no. 2, pp. 1980–1991, 2024, doi: 10.11591/ijai.v13.i2.pp1980-1991.

[3] A. D. Gibson, N. M. White, G. S. Collins, and A. G. Barnett, “Evidence of Unreliable Data and Poor Data Provenance in Clinical Prediction Model Research and Clinical Practice,” BMC Medicine, vol. 24, article 386, 2026, doi: 10.1186/s12916-026-04981-y.

[4] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons, “Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): The TRIPOD Statement,” BMJ, vol. 350, g7594, 2015, doi: 10.1136/bmj.g7594.

[5] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, and K. Crawford, “Datasheets for Datasets,” Communications of the ACM, vol. 64, no. 12, pp. 86–92, 2021, doi: 10.1145/3458723.

[6] R. Y. Wang and D. M. Strong, “Beyond Accuracy: What Data Quality Means to Data Consumers,” Journal of Management Information Systems, vol. 12, no. 4, pp. 5–33, 1996, doi: 10.1080/07421222.1996.11518099.

[7] 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, article 15, 2012, doi: 10.1145/2382577.2382579.

[8] S. Kapoor and A. Narayanan, “Leakage and the Reproducibility Crisis in Machine-Learning-Based Science,” Patterns, vol. 4, no. 9, article 100804, 2023, doi: 10.1016/j.patter.2023.100804.

[9] M. Baker, “1,500 Scientists Lift the Lid on Reproducibility,” Nature, vol. 533, no. 7604, pp. 452–454, 2016, doi: 10.1038/533452a.

[10] M. R. Munafò, B. A. Nosek, D. V. M. Bishop, K. S. Button, C. D. Chambers, N. Percie du Sert, U. Simonsohn, E.-J. Wagenmakers, J. J. Ware, and J. P. A. Ioannidis, “A Manifesto for Reproducible Science,” Nature Human Behaviour, vol. 1, article 0021, 2017, doi: 10.1038/s41562-016-0021.

[11] J. P. A. Ioannidis, “Why Most Published Research Findings Are False,” PLoS Medicine, vol. 2, no. 8, e124, 2005, doi: 10.1371/journal.pmed.0020124.

[12] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, “Model Cards for Model Reporting,” in Proc. Conf. on Fairness, Accountability, and Transparency (FAT* ’19), Atlanta, GA, USA, 2019, pp. 220–229, doi: 10.1145/3287560.3287596.

[13] I. P. Fellegi and A. B. Sunter, “A Theory for Record Linkage,” Journal of the American Statistical Association, vol. 64, no. 328, pp. 1183–1210, 1969, doi: 10.1080/01621459.1969.10501049.

[14] H. He and E. A. Garcia, “Learning from Imbalanced Data,” IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263–1284, 2009, doi: 10.1109/TKDE.2008.239.

[15] T. Saito and M. Rehmsmeier, “The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets,” PLOS ONE, vol. 10, no. 3, e0118432, 2015, doi: 10.1371/journal.pone.0118432.

[16] R. Kohavi, “A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection,” in Proc. 14th International Joint Conference on Artificial Intelligence (IJCAI), Montreal, Canada, vol. 2, 1995, pp. 1137–1143.

[17] S. Ahmad Bhat, “Lung Cancer,” Kaggle, 2021. [Online]. Available: https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer.

[18] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-Sampling Technique,” Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.

[19] IEEE, “IEEE Editorial Style Manual for Authors,” IEEE Editorial, 2019.

[20] O. Tur, A. Tur, V. Shabunina, and E. Chernaia, “The IEEE Style: Peculiarities of the Format and Application Prospects,” in Proc. 25th IEEE Int’l Conf. on Problems of Automated Electric Drive. Theory and Practice (PAEP), 2020, doi: 10.1109/PAEP49887.2020.9240790.