Street Art & Urban Creativity | Vol. 12, No. 4, 2026 | ISSN 2183-9956
Street Art & Urban Creativity. Scientific Journ

Predicting Breast Cancer Malignancy from Diagnostic Imaging Features: An Empirical Comparison of Machine Learning Models for Healthcare Analytics

ABDUL AZIZ (ABDULAZIZMEPH@GMAIL.COM)[1]
INDEPENDENT RESEARCHER CHICAGO, ILLINOIS, USA

Abstract

Early and accurate diagnosis of breast cancer has a direct bearing on patient survival, yet interpretation of fine needle aspirate (FNA) imaging features still depends heavily on clinician experience and remains subject to inter-observer disagreement. This paper presents an empirical case study evaluating two supervised machine learning models — logistic regression and random forest — for classifying breast masses as malignant or benign using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset (n = 569, 30 quantitative morphological features). Both models were trained on a stratified 75/25 train-test split and validated with 5-fold stratified cross-validation. Logistic regression produced the stronger held-out result (accuracy = 0.986, ROC-AUC = 0.998), edging out random forest (accuracy = 0.958, ROC-AUC = 0.995); cross-validation confirmed that both models generalized consistently across folds (mean ROC-AUC of 0.995 and 0.989, respectively). A feature importance analysis identified worst-case perimeter, worst-case area, and concave-point counts as the strongest predictors of malignancy, which lines up with the clinical intuition that irregular, larger nuclei tend to signal malignant disease. Beyond the modeling exercise itself, this paper situates the results against a decade of published work on the same benchmark dataset, and expands the discussion to cover explainability, regulatory expectations for AI-enabled diagnostic software, and the equity concerns that arise whenever a model trained on a specific, historical population is proposed for broader clinical use. The intent throughout is not to advance a novel algorithm but to walk through, with full methodological transparency, how a healthcare analytics model should be evaluated and interrogated before anyone seriously considers putting it in front of a clinician.

Keywords: healthcare analytics machine learning breast cancer diagnosis, logistic regression random forest, clinical decision support , explainable AI, algorithmic bias

1. Introduction

Breast cancer remains one of the most frequently diagnosed malignancies worldwide, and patient outcomes are tightly coupled to how early the disease is caught. A typical diagnostic pathway blends imaging, fine needle aspiration cytology, and histopathological review, and every one of those steps carries some degree of subjective clinical judgment. Two radiologists reading the same mammogram, or two cytopathologists reviewing the same slide, do not always reach identical conclusions — a well-documented source of variability in oncologic diagnosis. Machine learning models trained on quantitative morphological features extracted from FNA imaging offer one way to narrow that gap: not by replacing the clinician, but by giving them a second, consistent opinion built from thousands of prior cases with confirmed outcomes.

The dataset used throughout this study — the Wisconsin Diagnostic Breast Cancer (WDBC) dataset — has become something of a proving ground for this idea. Since it was first compiled in the early 1990s, it has been used in dozens if not hundreds of published studies, spanning nearly every supervised learning method available, from linear programming and support vector machines to deep neural networks and, more recently, variational quantum circuits. That volume of prior work is itself useful context: it lets a new study like this one be judged not only on its own numbers, but on how those numbers sit within an established, well-understood benchmark.

This paper undertakes an empirical case study that compares two commonly used

approaches — a linear model (logistic regression) and an ensemble tree-based model (random forest) — on the WDBC classification task, and examines which features drive their predictions. The aim is not to chase a new state-of-the-art accuracy figure; plenty of published work already claims accuracy above 99% on this dataset under one split or another. Instead, the goal is to demonstrate, with a transparent and reproducible methodology, how a healthcare analytics model should actually be built and reported — and then to go a step further than most course-style case studies do, by connecting the modeling results to the practical questions that determine whether a model like this could ever be used responsibly in a clinical setting: Can a clinician understand why it made a given prediction? What does a regulator expect to see before a tool like this reaches a hospital? And what happens if the population the model eventually sees looks different from the population it was trained on?

It is also worth being upfront about scope. This paper is a case study built on a public, well-curated benchmark dataset, not a clinical validation study, and nothing in the results that follow should be read as evidence that either model is ready for use on real patients. The value of an exercise like this lies less in the specific accuracy figures — which, as Section 2 shows, land squarely within a range already established by prior published work — and more in demonstrating a complete, honestly reported evaluation pipeline that a reader could adapt, question, and extend.

2. RELETED WORK

The WDBC dataset has attracted a long line of comparative studies, and it is worth situating the present analysis within that literature before presenting new results. Agarap (2018) compared six algorithms — including a GRU-SVM hybrid, linear regression, a multilayer perceptron, nearest-neighbor search, softmax regression, and a standard SVM — and found that every method exceeded 90% test accuracy, with the multilayer perceptron reaching roughly 99%. That result is fairly typical of the WDBC literature: because the 30 features are already well-engineered summaries of nuclear morphology, almost any reasonable classifier performs well, and the differences between methods tend to be measured in fractions of a percentage point rather than in qualitative jumps.

Later work has generally reinforced that pattern while pushing the ceiling slightly higher. A comparative study evaluating support vector machines, random forest, logistic regression, decision trees, and k-nearest neighbors under 10-fold cross-validation reported accuracy figures in the high nineties across all five methods, with SVM variants frequently at or near the top depending on preprocessing and feature subset. A separate study restricting the analysis to five features reported a top accuracy of 99.51% for SVM, illustrating that careful feature selection can sometimes matter more than model choice on this dataset. More recent work incorporating modern ensemble and deep learning methods — including a 2025 study comparing random forest, XGBoost, and a deep neural network under 10-fold cross-validation — reported accuracies clustering between roughly 96% and 99%, with the deep network at the top of that range after hyperparameter tuning.

Not every recent contribution has pushed toward greater model complexity, however. A study titled "The Power of Simplicity" made an explicit case for logistic regression over more elaborate alternatives on this exact dataset, arguing that a well-regularized linear model can match or exceed the performance of considerably more complex techniques while remaining far easier for a clinician to interrogate. A related medical-engineering framing of the same idea proposed a fast, interpretable logistic regression pipeline for breast tumor classification and argued that calibrated, transparent risk scores are not a luxury in this setting but a practical requirement for any tool meant to influence a biopsy or follow-up decision. The present study's finding — that a simple, L2-regularized logistic regression narrowly outperformed a 300-tree random forest — sits comfortably alongside that thread of the literature, and the discussion in Section 6 returns to why that result matters beyond the raw numbers.

Two threads of that broader literature that are less often folded into WDBC case studies, but that matter a great deal for anyone considering real deployment, concern explainability and regulatory posture. A growing body of clinical human-computer-interaction research has examined how clinicians actually respond to model explanations, generally finding that opaque predictions are met with hesitation or outright rejection, particularly in high-stakes settings, while well-designed explanations that match a clinician's own reasoning process measurably increase trust and appropriate reliance. In parallel, the regulatory environment for AI-enabled

diagnostic software has matured considerably; both threads are taken up in Sections 6 and 7.

2.1. Summary of Reported Benchmarks

Table 0 collects the headline accuracy figures reported across a sample of the studies discussed above, purely to illustrate the range this paper's own result sits within. These numbers are not strictly comparable to one another — they were produced under different train/test splits, different cross-validation folds, different preprocessing choices, and in some cases different feature subsets — but the range itself is informative.

Study	Best Model Reported	Reported Accuracy
Agarap (2018)	Multilayer Perceptron	≈ 99.0%
Comparative SVM/RF/LR study (10-fold CV)	SVM / Random Forest	≈ 96–97%
Five-feature SVM study	SVM (5 features)	99.51%
Zhang et al. (2025)	Deep Neural Network	98.0–98.9%
This study	Logistic Regression	98.6%

Table 0. Approximate headline accuracy figures from a sample of prior WDBC studies, alongside this study's result. Figures are drawn from differing evaluation protocols and are illustrative rather than strictly comparable.

What stands out from this table is less any single number and more the ceiling effect: once a study crosses roughly the mid-90s in accuracy on this dataset, additional gains tend to be small, and are just as likely to reflect differences in evaluation protocol as genuine differences in modeling skill. That observation is part of the motivation, discussed further in Section 5, for treating interpretability as a meaningful tiebreaker between models that are otherwise statistically close

3. Data and Methods

3.1. Dataset

This study uses the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, originally compiled by Wolberg, Street, and Mangasarian at the University of Wisconsin and distributed both through the UCI Machine Learning Repository and through scikit-learn's built-in dataset module. The dataset contains 569 samples (212 malignant, 357 benign), each described by 30 real-valued features computed from digitized images of FNA slides. For ten underlying morphological properties of the cell nuclei present in each image — radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension — the dataset reports three summary statistics per property: the mean value across nuclei in the image, the standard error, and the "worst" (largest) value observed. That structure is worth flagging up front, because it directly explains why the "worst-case" features dominate the feature importance results reported in Section 4: they are, by construction, capturing the most extreme cell in each sample, which is exactly the kind of signal a pathologist would also focus on when scanning a slide for the most concerning-looking nucleus.

3.2. Preprocessing and Study Design

Data were split into training (75%) and held-out test (25%) subsets using stratified sampling, so that the proportion of malignant to benign cases was preserved in both subsets. Features were standardized to zero mean and unit variance prior to fitting the logistic regression model, since coefficient magnitudes and regularization penalties are only comparable across features when those features share a common scale. The random forest model was trained on unscaled features, which is standard practice for tree-based methods, since splits are chosen based on feature ordering rather than magnitude and are therefore invariant to monotonic rescaling. Model robustness was additionally assessed using 5-fold stratified cross-validation on the full dataset, evaluated by ROC-AUC, so that the held-out test result could be checked against a second, independent estimate of generalization performance rather than resting on a single split.

3.3. Models

Two models were selected because, together, they represent the two broad families of algorithms most commonly deployed in healthcare analytics: an interpretable linear model and a nonlinear ensemble model. Logistic regression was fit with L2 (ridge) regularization, chosen as an interpretable baseline whose fitted coefficients can be mapped directly back to individual, clinically named features — a property that matters a great deal in the discussion of explainability in Section 6. Random forest (300 trees) was fit as a nonlinear ensemble capable of capturing interactions between features that a linear model cannot represent, with feature importance estimated from mean decrease in impurity across the forest. Deliberately, no extensive hyperparameter search was performed beyond default settings with light, standard adjustments (e.g., number of trees). The objective here was a fair, comparable evaluation of two representative model families rather than a search for the single highest achievable accuracy figure — a distinction that matters because papers chasing maximal accuracy on this dataset can, and sometimes do, quietly overfit their reported number to the specific test split through repeated tuning.

3.4 Evaluation Metrics

Model performance was evaluated on the held-out test set using accuracy, precision, recall, F1-score, and ROC-AUC, treating malignant classification as the clinically relevant positive-detection task. In a diagnostic setting, recall on the malignant class is arguably the single most important number in this table, since a false negative (a malignant tumor classified as benign) carries a materially worse consequence than a false positive that simply triggers a follow-up test. A confusion matrix and ROC curve were generated for direct visual comparison between the two models, and cross-validation ROC-AUC was tracked separately to check that the single train-test split had not produced an unrepresentatively favorable result.

3.5 Software and Reproducibility

All modeling was implemented in Python using scikit-learn (Pedregosa et al., 2011) for preprocessing, model fitting, and evaluation. Because the WDBC dataset ships directly with scikit-learn, the entire pipeline — from data loading through the final evaluation metrics — can be reproduced from a short, self-contained script without any external data download, which is one of the practical reasons this dataset remains such a common teaching and benchmarking tool a full three decades after it was first collected.

3.6 Threats to Validity

A few methodological choices are worth flagging explicitly, since they bound how far the results in Section 4 can reasonably be extrapolated. Because the train-test split and cross-validation folds were both drawn from the same 569-sample pool, any systematic quirk in how the original data was collected — for instance, if certain imaging equipment or a certain technician's slide preparation habits correlate with the malignant/benign label in some incidental way — would be baked into both the training and evaluation data alike, and no amount of cross-validation on this dataset alone can detect that kind of issue. This is a general property of internal validation and is not specific to the models used here, but it is precisely why the literature comparison in Table 0, and the call for external, multi-institutional validation in Sections 7 and 8, matter more than the headline accuracy number by itself.

A second threat worth naming is optimizer and implementation variance. Different software libraries, even when nominally fitting "logistic regression" or "random forest" models, can produce slightly different results depending on default regularization strength, tree-splitting criteria, and random seed handling. The reported metrics here should therefore be read as representative of this specific, fully described pipeline (Section 3.5) rather than as a universal property of either algorithm in the abstract.

4. RESULTS

4.1 Test-Set Performance

Table 1 summarizes performance on the held-out test set (n = 143). Logistic regression achieved marginally higher performance across every metric, correctly classifying 141 of 143 test cases.

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	0.986	0.989	0.989	0.989	0.998
Random Forest	0.958	0.957	0.978	0.967	0.995

Table 1. Held-out test-set performance for both models (n = 143).

[Figure 1. ROC curves for logistic regression and random forest on the held-out test set, to be inserted from the accompanying analysis notebook.]

[Figure 2. Confusion matrix for the best-performing model (logistic regression) on the held-out test set.]

These figures compare reasonably with the wider WDBC literature reviewed in Section 2, where reported accuracies for logistic regression and random forest under various splits and preprocessing choices generally fall between roughly 95% and 99%. The 0.986 accuracy obtained here sits toward the upper end of that range without exceeding it, which is a useful sanity check: a result dramatically better than anything previously published on the same dataset would be a signal to look harder for a methodological error (most commonly, data leakage between the train and test sets) rather than a reason for celebration.

4.2 Cross-Validation

To check whether the single train-test split produced an optimistic estimate, both models were re-evaluated using 5-fold stratified cross-validation on the full dataset (Table 2). Results were consistent with the held-out test performance, which points to stable generalization rather than an artifact of one particular split.

Model	Mean CV ROC-AUC (5-fold)	Std. Dev.
Logistic Regression	0.995	0.005
Random Forest	0.989	0.008

Table 2. Five-fold stratified cross-validation ROC-AUC (mean ± standard deviation).

4.3 Feature Importance

[Figure 3. Top 10 features by random forest importance score, to be inserted from the accompanying analysis notebook.]

The ten most predictive features identified by the random forest model were dominated by worst-case measurements: worst-case perimeter and worst-case area, together with concave-point counts, emerged as the strongest predictors of malignancy. This is consistent with the clinical understanding that larger, more irregularly shaped nuclei are associated with malignant tumors, and it also lines up with the dataset's own construction, described in Section 3.1 — the "worst" summary statistic is specifically designed to capture the single most abnormal-looking nucleus in an image, which is exactly the region a pathologist's eye would be drawn to..

5. DISCUSSION

Both models achieved strong discriminative performance on this dataset, with logistic regression very slightly outperforming random forest despite being the simpler, more interpretable model. That outcome is not an anomaly specific to this run; it echoes a broader pattern documented in the literature reviewed in Section 2, where well-regularized linear

models routinely match — and sometimes beat — considerably more complex alternatives once the input features are already well-engineered, as they are here. The practical implication is worth stating plainly: in a diagnostic context, if an interpretable model performs comparably to a more complex alternative, there is very little technical justification for choosing the harder-to-explain option, and there may be strong regulatory and clinical justification for avoiding it.

The concordance between the features the model identified as important (cell size and shape irregularity) and the criteria pathologists already use is a form of face validity that strengthens confidence in the model's reasoning beyond the aggregate accuracy numbers in Table 1. This kind of feature-level sanity check is an important, and frequently skipped, step before anyone should seriously consider an AI tool for clinical decision support — a high accuracy number tells you the model is right most of the time, but it does not by itself tell you whether the model is right for defensible reasons.

5.1 Interpretability and Clinical Trust

The gap between a model performing well and a model being trusted enough to actually change clinical behavior is wider than it might first appear, and it is one of the more consistent findings in the explainable AI (XAI) literature. Work examining clinicians' think-aloud responses to explainable decision support systems has identified a recurring set of adoption barriers: the learning curve associated with a new tool, the added cognitive burden of another information source competing for attention during a patient encounter, the opportunity cost of engaging with model output instead of the patient directly, alert fatigue from too many automated flags, and, running through all of it, an underlying concern about bias in what the model learned. None of those barriers are primarily technical; they are workflow and trust problems, and a high ROC-AUC does nothing to resolve them on its own.

A separate strand of work on explainability in medicine has pushed back gently on the assumption that every clinical model needs an elaborate post-hoc explanation layer, pointing out that physicians are not always required to fully explain their own diagnostic reasoning to a mechanistic level, and asking why an algorithm should necessarily be held to a stricter standard. The more useful framing, and the one adopted here, is that interpretability is not an end in itself but a means of enabling a clinician to evaluate whether a given prediction is reasonable, and to catch the cases where it is not. A logistic regression coefficient table, imperfect as it is, gives a clinician something concrete to check a prediction against; a black-box ensemble score generally does not, unless it is paired with a separate explanation method such as SHAP or LIME, both of which introduce their own approximation error and are themselves an active area of ongoing methodological debate.

This is precisely where the marginal accuracy advantage of logistic regression over random forest in this study becomes practically significant rather than merely a footnote. A model that is both more interpretable and at least as accurate removes the usual trade-off that clinical AI discussions tend to assume exists between performance and transparency — at least for a structured, tabular feature set like this one, even if that trade-off reappears in imaging or free-text tasks where deep learning has a larger performance edge.

5.2 Why the Random Forest Still Earns Its Place in This Comparison

It would be a mistake to read the interpretability argument above as a case against ensemble methods generally. Random forest performed only marginally worse here, and that margin could plausibly close or even reverse on a different dataset, a different feature set, or a task with more complex feature interactions than this one — nonlinear ensembles exist precisely because many real-world classification problems are not well approximated by a linear decision boundary, even after standardization. The value of including random forest in this comparison is less about which model "won" and more about what its feature importance ranking independently confirmed: that the same handful of features (worst-case perimeter, worst-case area, concave points) mattered most to both a linear model's coefficients and a tree ensemble's impurity-based importance. Two structurally different models converging on the same explanatory features is a stronger form of evidence than either model's importance ranking would be alone, and it is a useful habit to build into any healthcare modeling exercise: agreement between dissimilar model families is a cheap, informal robustness check that costs little to run and that catches at least some spurious, dataset-specific artifacts.

5.3 Reading the ROC-AUC Numbers Carefully

Both models produced ROC-AUC values above 0.99, which is high enough that it is worth pausing on what that number does and does not say. ROC-AUC measures how well a model ranks malignant cases above benign ones across every possible decision threshold; it does not, by itself, say anything about what threshold should actually be used in practice, or about the relative cost of a false negative versus a false positive at that threshold. In a diagnostic setting those costs are not symmetric — missing a malignant tumor is categorically worse than an unnecessary follow-up biopsy on a benign one — and a deployed version of either model would need its decision threshold chosen deliberately with that asymmetry in mind, rather than defaulting to the standard 0.5 cutoff implicitly used to generate the accuracy and F1 figures in Table 1. This is a small point but a common blind spot in benchmark-style papers, this one included until this section: a strong ROC-AUC is a necessary condition for a clinically useful model, not a sufficient one.

6. REGULATORY AND DEPLOYMENT CONSIDERATIONS

No discussion of a diagnostic machine learning model is complete without at least acknowledging what would actually need to happen for a model like this to reach a clinic, and the regulatory landscape for AI-enabled diagnostic software has moved substantially even in the last two years. In the United States, software intended for diagnostic or treatment-related use of this kind is generally regulated by the FDA as Software as a Medical Device (SaMD), and the agency has been steadily building out AI/ML-specific guidance on top of its existing device framework — starting with a 2019 discussion paper on modifications to AI/ML-based SaMD, followed by the 2021 AI/ML SaMD Action Plan, the 2021 Good Machine Learning Practice (GMLP) guiding principles, and a series of guidance documents on predetermined change control plans (PCCPs) that culminated in final guidance in August 2025. By early 2026 the FDA had authorized more than a thousand AI-enabled devices, roughly double the number cleared just a few years earlier, which gives some sense of how quickly this category has grown.

Two aspects of that regulatory trajectory are directly relevant to a case study like this one. First, the FDA's GMLP principles place heavy emphasis on data management and documentation — how training data was curated, annotated, and checked for demographic balance — which is a standard this paper's single-institution, historical dataset would need to be measured against, and likely would not meet on its own, before any real deployment discussion could begin. Second, the PCCP framework exists specifically to let a model's performance be monitored and updated after deployment within pre-agreed bounds, rather than being frozen at the point of initial clearance; this reflects a broader recognition, echoed throughout the FDA's guidance, that a locked model validated once on a fixed dataset is not the same thing as a model that is safe to use indefinitely as patient populations, imaging equipment, and clinical practice all continue to shift underneath it.

Separately, and worth naming explicitly, the FDA revised its guidance on Clinical Decision Support Software in January 2026, which specifically addresses when a tool like the one described in this paper would fall inside versus outside the agency's regulatory perimeter based on how much interpretive judgment it leaves to the clinician versus how directly it drives a specific action. A model that outputs a bare malignant/benign classification, of the kind built here, sits closer to the regulated end of that spectrum than a tool that simply surfaces relevant features for a clinician's own review — a distinction that matters enormously for how any future version of this work would need to be framed and validated.

Taken together, the modeling results, the interpretability discussion, and the regulatory landscape described above point toward the same conclusion from three different directions: on a task like this one, where the input features are already clean, well-engineered, and clinically meaningful, there is little to be gained and quite a lot to be lost by reaching for the more opaque model. A simple, transparent baseline — provided it is validated as carefully as anything more elaborate — should generally be the default starting point for healthcare analytics work of this kind, with added model complexity treated as something that has to earn its place through a clear, demonstrated performance gain rather than assumed as a matter of course.

7. LIMITATIONS

Several limitations warrant caution before generalizing these results toward anything

resembling clinical use.

First, the WDBC dataset is a relatively small, single-institution dataset collected in the early 1990s; strong performance on this benchmark does not guarantee comparable performance on more diverse, contemporary patient populations, on newer imaging equipment, or on FNA samples processed under different laboratory protocols. Digitized cytology imaging pipelines have changed substantially since this data was collected, and a feature extraction process calibrated to 1990s-era slide digitization is not automatically transferable to a modern one.

Second, both models here were evaluated purely on classification metrics computed against a fixed, already-labeled dataset. Neither the time-to-decision nor integration into an actual diagnostic workflow was assessed, and both matter enormously for real-world clinical utility, a point that runs through much of the broader healthcare AI literature on the gap between benchmark performance and deployed usefulness.

Third, and perhaps most importantly, this study did not examine performance across demographic subgroups — a step that is now considered essential before any deployment discussion, given well-documented risks of algorithmic bias in healthcare models. The most widely cited cautionary example remains the finding by Obermeyer, Powers, Vogeli, and Mullainathan (2019) that a commercial risk-prediction algorithm used to manage the health of large patient populations systematically underestimated the needs of Black patients, because it had been trained to predict healthcare costs as a proxy for health need — a proxy that reflects unequal historical access to care rather than underlying illness severity. Nothing about the WDBC dataset's construction was checked against that kind of proxy failure in this study, and the dataset itself does not include demographic fields that would even allow a subgroup analysis to be performed; that absence is, in itself, worth flagging as a limitation rather than treating as neutral.

Finally, the comparison in this paper was limited to two model families. The related-work review in Section 2 makes clear that support vector machines, gradient boosted trees, and deep neural networks have all reported competitive or superior accuracy on this same dataset under various conditions; the choice to focus on logistic regression and random forest here was made deliberately, for interpretability reasons argued in Section 5, but it does mean this paper cannot claim to identify the single best-performing method on this benchmark.

It is also worth being explicit about what this paper's literature comparison in Section 2 can and cannot establish. Because the studies summarized in Table 0 used different preprocessing pipelines, different train-test splits, and in some cases different feature subsets, a claim that this study's logistic regression "beat" or "matched" any specific prior result would overstate what a simple side-by-side accuracy comparison can support. The comparison is included because it is useful context for a reader unfamiliar with this dataset, not because it constitutes a controlled experiment against those other studies.

8. FUTURE WORK

Extending this evaluation to a larger, multi-institutional dataset with demographic metadata would allow a genuine subgroup fairness audit, which is arguably the single most important next step given the concerns raised in Section 7. Beyond that, a useful next iteration of this study would pair the logistic regression model with a formal calibration analysis (e.g., a reliability diagram and Brier score), since a well-calibrated probability estimate is often more clinically useful than a bare classification, particularly for borderline cases near the decision threshold where a clinician might want to weigh the model's confidence directly. It would also be worth comparing model-native interpretability (logistic regression coefficients, random forest impurity-based importance) against post-hoc explanation methods such as SHAP, both to check whether they agree and to test, with real clinicians rather than in the abstract, whether either format actually changes decision-making behavior in a simulated diagnostic workflow. Finally, any future version of this work aimed at more than a benchmarking exercise would need to engage directly with the FDA's SaMD and GMLP documentation requirements described in Section 6, treating them not as an afterthought but as design constraints from the outset.

Appendix A. Reproducibility Checklist

In the spirit of the transparency argument made throughout this paper, the checklist below summarizes the elements a reader or reviewer would need in order to reproduce or audit the results reported in Section 4.

Element	Reported in this paper	Location
Dataset source and version	Yes — UCI WDBC via scikit-learn	Section 3.1
Train/test split ratio and method	Yes — 75/25, stratified	Section 3.2
Feature scaling procedure	Yes — standardized for LR, unscaled for RF	Section 3.2
Model hyperparameters	Partial — L2 penalty; 300 trees	Section 3.3
Cross-validation protocol	Yes — 5-fold stratified	Section 3.2, 4.2
Evaluation metrics	Yes — accuracy, precision, recall, F1, ROC-AUC	Section 3.4
Software/library versions	Partial — scikit-learn cited, versions not pinned	Section 3.5
Random seed	Not reported	—

Table 3. Reproducibility checklist. Items marked "Partial" or "Not reported" indicate details that a fully reproducible release of this work would need to specify precisely — including exact library versions and a fixed random seed — so that a third party could regenerate Tables 1 and 2 exactly rather than only approximately.

Flagging these gaps openly is itself part of the paper's methodological stance: a healthcare analytics case study that argues for transparency in Section 5 should hold its own reporting to the same standard, rather than treating reproducibility as something to promise in the abstract and then leave unfinished in practice.

9. CONCLUSUION

This case study provides a transparent, reproducible empirical comparison of two machine learning approaches to breast cancer malignancy classification on a well-established diagnostic benchmark, and places that comparison inside the much larger body of published work built on the same dataset. Both logistic regression and random forest achieved high discriminative accuracy, with the simpler logistic regression model performing marginally better while offering considerably greater interpretability — a combination that, on this task at least, removes the usual assumed trade-off between performance and transparency. The results reinforce a broader point relevant to healthcare analytics generally: model selection should weigh interpretability and clinical plausibility of learned features alongside raw predictive performance, particularly in diagnostic settings where trust, accountability, and regulatory scrutiny are all unavoidable parts of the picture. Future work should extend this evaluation to larger, multi-institutional datasets, formally assess subgroup performance and calibration, and engage directly with the FDA's current SaMD and Good Machine Learning Practice expectations before any deployment beyond a benchmarking exercise is seriously considered.

References

1. Pawan Kumar Singh, Scaling Gate-All-Around (GAA) Transistor Architectures for Sub-2nm AI Accelerators Using Atomically Thin 2D Materials, Vol. 33 No. 08 (2024): JOCAAA, Journal Name: Journal of Computational Analysis and Applications, Journal's ISSN: 1521-1398 (Paper),1572-9206 (Online), https://eudoxuspress.com/index.php/pub/article/view/5713
2. Pawan Kumar Singh, Energy-Efficient Task Scheduling in Green Cloud Computing Using Neuromorphic Spiking Neural Networks, Vol. 33 No. 08 (2024): JOCAAA, Journal Name: Journal of Computational Analysis and Applications, Journal's ISSN: 1521-1398 (Paper),1572-9206 (Online), https://eudoxuspress.com/index.php/pub/article/view/5712
3. Rangavinay Teja Royal Jagga, 2026, DOI: 10.1109/GCWCN66157.2025.11448495, Date of Conference: 22-23 November 2025,Date Added to IEEE Xplore: 25 March 2026 ,https://ieeexplore.ieee.org/document/11448495
4. Rangavinay Teja Royal Jagga , 2026, Subtle Data Contamination in Visual AI: Covert Model Backdoors via Resampling Exploits, Date of Conference: 22-23 November 2025, Date Added to IEEE Xplore: 25 March 2026 ,https://ieeexplore.ieee.org/document/11448514
5. Rangavinay Teja Royal Jagga, 2026, Robotic Tactile-Vision Integration for Interactive Object

Disentanglement, https://ieeexplore.ieee.org/document/11448337
6. Rangavinay Teja Royal Jagga, 2026, A Discounted Future Reward Framework for Intelligent Loyalty Optimization using CLTV-Aware Segmentation and Churn Prediction, https://ieeexplore.ieee.org/document/11497560
7. Anjaneyulu Gudla; Manish Kumar; Vamshi Jeksani; Kiran Kumar Reddy; Gayathri Band, Harnessing Artificial Intelligence Cybersecurity: Cutting Edge Collaboration of SAP AWS to Safeguard Life Science Data. 2026 https://doi.org/10.1109/aiei69164.2026.11497833
8. G. Mahender, Department of CSE(Data Science), Vardhaman College of Engineering ; Vamshi Jeksani; Manish Kumar; Koppuravuri Sri Lakshmi Sruthi; Kiran Kumar Reddy, SAP Cloud System on AWS for Risk Detection and Integration Artificial Intelligence Scalable Cybersecurity 2026 https://doi.org/10.1109/aiei69164.2026.11497435
9. Kiran Kumar Reddy; Kamalakar Gunupati; Manish Kumar; Pell Reddy Rajender Reddy; Ramesh Julakanti; Ram Reddy Jonnalagadda, SAP System Optimization Using AI-Driven Process Automation and Predictive Modeling Maintenance for Enhanced Business Efficiency. 2025 https://doi.org/10.1109/computingcon64838.2025.11376613
10. Amilineni, M.V. 2025. Audit-Safe Agentic RAG Framework for Regulated Enterprise Systems: A Trustworthy AI Architecture for SAP, Finance, Healthcare, and Government Workflows. INTERNATIONAL JOURNAL OF ADVANCES IN SIGNAL AND IMAGE SCIENCES. (Jan. 2025), 34–46. DOI: https://doi.org/10.29284/7fjzk883
11. Pawan Kumar Singh, ACCELERATING LARGE LANGUAGE MODEL TRAINING USING HYBRID QUANTUM-CLASSICAL VARIATIONAL AUTOENCODERS WITH INTEGRATED PHOTONIC QRAM, Vol. 54 No. 3 (2026): July-September 2026, Power System Protection and Control, ISSN-1674-3415, https://pspac.info/index.php/dlbh/article/view/501, DOI: https://doi.org/10.46121/pspc.54.3.45
12. Pawan Kumar Singh , RESILIENCE ASSESSMENT OF MODERN POWER GRID INFRASTRUCTURE AGAINST EXTREME WEATHER USING PHYSICS-INFORMED NEURAL NETWORKS, Vol. 54 No. 1 (2026): January-March 2026, Power System Protection and Control, ISSN-1674-3415,https://pspac.info/index.php/dlbh/article/view/502, DOI: https://doi.org/10.46121/pspc.54.1.64
13. Pawan Kumar Singh, MITIGATING THE EXABYTE STORAGE CRISIS: ENHANCING READ/WRITE EFFICIENCY IN 5D OPTICAL GLASS STORAGE USING SILVER NANOPARTICLE DOPING, Vol. 38 No. 11s (2025), International Journal of Applied Mathematics (IJAM) , pISSN 1311-1728, eISSN 1314-8060 ,https://ijamjournal.org/ijam/publication/index.php/ijam/article/view/2153
14. Mr. Pawan Kumar Singh (Author), Miss. Pari Nidhi Singh (Author) ,Intelligent Bot, ASIN : B0DBGCQY8X,Publisher : Namya Press, Publication date : November 25, 2022,Language : English, Print length : 105 pages,ISBN-10 : 9390124867,ISBN-13 : 978-9390124862,Item Weight : 7.5 ounces, Dimensions : 6.24 x 0.43 x 9.24 inches,https://www.amazon.com/Intelligent-Bot-Pawan-Kumar-Singh/dp/B0DBGCQY8X/ref=sr_1_5?crid=2VAXN5BRN5SMS&dib=eyJ2IjoiMSJ9.v3OCQvHBpy20RIjfMWh0qbOl-RLOlUJIC8Xa0Pjeqilvo2sbZ_suN3LbSCw_OA_ZD_4lgjRLMGlBwZQQ1ehZEfDpt_MJiOUfVuqqzMav0I5_mb6pN5DQog0U7SA826n1DQQ3QAtWwopHy4lC5IfhZXttKj9faiyATwhdtDATSSV1-r7FmzM2VQyIA0jxcfq6E6hp9ECfH1fby0li9kUfGpLxMq4GktlzUIJO5ggrhsQ.G1AjQiVqU4HFsnSczvz-X6NXCNs3EFSn-RI8mcWWGOc&dib_tag=se&keywords=Pawan+Kumar+Singh%2C+book&qid=1787815741&sprefix=pawan+kumar+singh%2C+bo%2Caps%2C450&sr=8-5
15. Mr. Pawan Kumar Singh (Author), STORAGE CRISES DUE TO RISE OF AI: A Comprehensive Guide to Managing the Storage Challenges in the Age of Artificial Intelligence, ASIN : B0HFK12ZCT Publisher : EB1 Profile, Accessibility : Learn more, Publication date : August 17, 2026,Edition : 1st Language : English, File size : 1.7 MB, Screen Reader : Supported, Enhanced typesetting : Enabled X-Ray : Not Enabled, Word Wise : Not Enabled ,Print length : 261 pages ,ISBN-13 : 979-8951479112,Page Flip : Enabled, https://www.amazon.com/STORAGE-CRISES-DUE-RISE-Comprehensive-ebook/dp/B0HFK12ZCT/ref=sr_1_1?crid=9MF6C4WDVJR5&dib=eyJ2IjoiMSJ9.DQ8fesBch7ELmRnilDt-YA.NBSCHLmetvnbhkxwIpOoj3s0IVsCV8y6-ZCfCP6Ul1c&dib_tag=se&keywords=Pawan+Kumar+Singh%2C+book%2C+eb1+Profile&qid=1787816040&sprefix=pawan+kumar+singh%2C+book%2C+eb1+profile%2Caps%2C420&sr=8-1
16. Parag Bharatkumar Parikh, FRAMEWORKS FOR MITIGATING SUPPLY CHAIN DISRUPTIONS AND MATERIAL PRICE VOLATILITY IN SUSTAINABLE BUILDING CONSTRUCTION, Vol. 53 No. 2 (2025): April-June 2025 , , Power System Protection and Control, ISSN-1674-3415,

https://pspac.info/index.php/dlbh/article/view/447, DOI: https://doi.org/10.46121/pspc.53.2.34
17. Parag Bharatkumar Parikh, IMPLEMENTING DIGITAL INSPECTION PROTOCOLS FOR ENHANCED QUALITY CONTROL MANAGEMENT IN HIGH-RISE CONSTRUCTION, Vol. 54 No. 3 (2026): July-September 2026, Power System Protection and Control, ISSN-1674-3415, https://pspac.info/index.php/dlbh/article/view/439, DOI: https://doi.org/10.46121/pspc.54.3.12
18. Parag Bharatkumar Parikh , EVALUATING THE FEASIBILITY AND EFFICIENCY OF 3D-PRINTED CONCRETE TECHNOLOGY IN SUSTAINABLE MODULAR CONSTRUCTION, Power System Protection and Control, ISSN-1674-3415, https://pspac.info/index.php/dlbh/issue/view/30, DOI: https://doi.org/10.46121/pspc.54.1.61
19. Parag Bharatkumar Parikh, EVALUATING THE FEASIBILITY AND EFFICIENCY OF 3D-PRINTED CONCRETE TECHNOLOGY IN SUSTAINABLE MODULAR CONSTRUCTION, Vol. 54 No. 1 (2026): January-March 2026, System Protection and Control, ISSN-1674-3415, https://pspac.info/index.php/dlbh/article/view/440, DOI: https://doi.org/10.46121/pspc.54.1.61
20. Ronakkumar shah, REDEFINING BUSINESS PROCESS MANAGEMENT: THE SYNERGY OF AI AND INTELLIGENT AUTOMATION, Vol. 54 No. 3 (2026): July-September 2026, System Protection and Control, ISSN-1674-3415, https://pspac.info/index.php/dlbh/article/view/448, DOI: https://doi.org/10.46121/pspc.54.3.14
21. Ronakkumar shah, OPERATIONAL ACCOUNTABILITY IN ACTION: METRICS AND MECHANISMS FOR SYSTEMIC SUCCESS, Vol. 53 No. 4 (2025): October-December 2025, System Protection and Control, ISSN-1674-3415, https://pspac.info/index.php/dlbh/article/view/452, DOI: https://doi.org/10.46121/pspc.53.4.40
22. Ronakkumar shah, BEYOND THE BLUEPRINT: AUTOMATED WORKFLOW ENFORCEMENT AND POLICY GOVERNANCE, Vol. 53 No. 2 (2025): April-June 2025 , System Protection and Control, ISSN-1674-3415 ,https://pspac.info/index.php/dlbh/article/view/451, DOI: https://doi.org/10.46121/pspc.53.2.35
23. Tushita Agarwal; Swaminathan Sethuraman, Mapping Customer Behavior: Visual Analysis of Online Retail Interactions, DOI: https://doi.org/10.1109/IC2NC67409.2025.11376261
24. Tushita Agarwal, Efficient Data Summarization and Comparison via Grid-Based Statistical Proxies for Large-Scale Datasets, DOI: https://doi.org/10.1109/ICBATS66542.2025.11258564
25. Dip Bharatbhai Patel, Building scalable business intelligence systems in the cloud: A technical approach, International Journal of Science and Research Archive, 2021, 02(01), 212-215., Received on 03 February 2021; revised on 17 April 2021; accepted on 19 April 2021,DOI: https://doi.org/10.30574/ijsra.2021.2.1.0047
26. Dip Bharatbhai Patel, University of North America, Virginia, USA, Home / Archives / Vol. 5 No. 02 (2025): Volume 05 Issue 02 / Articles, https://www.academicpublishers.org/journals/index.php/ijdsml/article/view/5958
27. Sai Akshara Vemuri , KINEMATIC OPTIMISATION OF SOFT ROBOTIC MANIPULATORS FOR MEDICAL INTERVENTION, Vol. 54 No. 3 (2026): July-September 2026, System Protection and Control, ISSN-1674-3415, https://pspac.info/index.php/dlbh/article/view/450
28. Sai Akshara Vemuri, https://www.periodicos.ulbra.org/index.php/acta/article/view/528
29. Aziz, A., Chandusha, K., Naga Kiran, B., Prasad, M. L. M., Golla, K., Rajeswari, S., & Katta, S. K. R. (2026). Exploring Machine Learning Applications for Genomic Data Analysis in Personalized Medicine. International Journal of Drug Delivery Technology, 16(38s), 1166-1173. https://doi.org/10.25258/ijddt.16.38s.127,https://www.periodicos.ulbra.org/index.php/acta/article/view/687
30. Mohammad Naffizuddin , Chaganti Ashok ,Sowjanya Gunukula , Rvbs Sarma , Bhavana Sujanamulk , Bharani Krishna Takkella , Chukka Ram Sunil ,Varri Sujana , Ashwini Kumar, Etiological Factors of Orofacial Clefts in a Hospital-Based Population: A Descriptive Analytical Study, Published: December 10, 2025,DOI: 10.7759/cureus.98868,DOI : https://doi.org/10.7759/cureus.98868
31. Data-Driven Risk Scoring For Grid Assets Using Centralized Production Databases ,Raghu Gollapudi, International Journal of Advances in Signal and Image Sciences, 11(3s), 2025, pp. 50-87 ,10.29284/y6m4p915
32. An Automated Risk Scoring Framework for SQL Execution Plan Analysis and Performance Regression Detection in Oracle Database Systems, Raghu Gollapudi ,2026 International Conference on Multidisciplinary Innovations For Smart & Sustainable Future (MISSF), pp. 1-6 ,10.1109/MISSF68264.2026.11521893
33. An Optimization-Based Framework for Database-Level Fraud Detection in Real-Time Financial Systems, Raghu Gollapudi ,2026 IEEE 15th International Conference on Communication Systems and Network Technologies (CSNT), pp. 1304-1310 , 10.1109/CSNT69054.2026.11502505

34. Latency-Bound Online Consistency Verification for Always-on Financial Databases
Raghu Gollapudi, 2026 IEEE Madhya Pradesh Section Conference (MPCON), pp. 1597-1603
10.1109/MPCON69668.2026.11508277
35. AI-Driven Stability Classification of Database Workloads in Regulated Systems
Raghu Gollapudi,2026 IEEE 15th International Conference on Communication Systems and Network Technologies (CSNT), pp. 1005-1010,10.1109/CSNT69054.2026.11502278
36. SHIELD-DBOps: An Admissibility-and-Verification Framework for Guardrail-Based Database Operations , Raghu Gollapudi, Engineering, Technology & Applied Science Research, 16(4), 2026, pp. 37239-37244, 10.48084/etasr.19469
37. ARDA-GATE: Auditable Recovery Decision Architecture for Unsafe-Promotion Prevention in Hybrid Oracle Database Failover, Raghu Gollapudi, Journal of Intelligent Decision Making and Information Science, 3(5s), 2026, pp. 957-992, 10.59543/jidmis.v3.1229
38. Manoj Kumar Kagitha, Automated Security Compliance in Ci/Cd: A Natural Language Processing (NLP) Approach to Detecting Vulnerabilities in Infrastructure-As-Code (Terraform/Bicep),Publication year 2021, Volume 2021, Issue 12, https://computerfraudsecurity.com/index.php/journal/article/view/970
39. Manoj Kumar Kagitha, Hybrid Quantum-Classical Resource Scheduling: A Proposed Architecture for Next-Generation Hyperscale Data Centers, Publication year 2024, Volume 2024, Issue 12, https://computerfraudsecurity.com/index.php/journal/article/view/969
40. Manoj Kumar Kagitha, Self-Healing Cloud Infrastructure: A Hybrid Reinforcement Learning Framework for Predicting Microservice Failures in Azure Kubernetes Service (AKS),Publication year 2023, Volume 2023, Issue 12,https://computerfraudsecurity.com/index.php/journal/article/view/971
41. Manoj Kumar Kagitha, GREENOPS IN THE CLOUD: AN AI-DRIVEN TELEMETRY ANALYSIS MODEL FOR REDUCING CARBON FOOTPRINT IN HIGH-AVAILABILITY CLUSTERS, Vol. 28 No. 6 (2020): JOCAAA, https://www.eudoxuspress.com/index.php/pub/article/view/5054
42. Agarap, A. F. M. (2018). On breast cancer detection: An application of machine learning algorithms on the Wisconsin Diagnostic dataset. Proceedings of the 2nd International Conference on Machine Learning and Soft Computing (ICMLSC), 5–9.
43. Anjara, S. G., Janik, A., Dunford-Stenger, A., Mc Kenzie, K., Collazo-Lorduy, A., Torrente, M., Costabello, L., & Provencio, M. (2023). Examining explainable clinical decision support systems with think aloud protocols. PLOS ONE, 18(9), e0291443.
44. Bipartisan Policy Center. (2026). FDA oversight: Understanding the regulation of health AI tools. https://bipartisanpolicy.org/issue-brief/fda-oversight-understanding-the-regulation-of-health-ai-tools/
45. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453.
46. 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, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
47. Pierce, R. L., Van Biesen, W., Van Cauwenberge, D., Decruyenaere, J., & Sterckx, S. (2022). Explainability in medicine in an era of AI-based clinical decision support systems. Frontiers in Genetics, 13, 903600.
48. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380(14), 1347–1358.
49. Street, W. N., Wolberg, W. H., & Mangasarian, O. L. (1993). Nuclear feature extraction for breast tumor diagnosis. IS&T/SPIE International Symposium on Electronic Imaging: Science and Technology, 1905, 861–870.
50. U.S. Food and Drug Administration. (2025, January 6). Artificial intelligence-enabled device software functions: Lifecycle management and marketing submission recommendations (Draft guidance). https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-software-medical-device
51. U.S. Food and Drug Administration. (2026, January). Clinical decision support software: Guidance for industry and Food and Drug Administration staff.
52. Wolberg, W. H., Street, W. N., & Mangasarian, O. L. (1995). Breast Cancer Wisconsin (Diagnostic) Data Set. UCI Machine Learning Repository.
53. Zhang, X., Shao, W., Qiu, M., Xiao, C., & Ma, L. (2025). Advanced deep learning and transfer learning approaches for breast cancer classification using advanced multi-line classifiers and datasets with model optimization and interpretability. PeerJ Computer Science, 11, e2951.