

INTERPRETABLE ENSEMBLE LEARNING FOR BREAST CANCER PREDICTION: FROM CYTOLOGICAL FEATURES TO CLINICAL DECISION SUPPORT: A REVIEW

Rafael Visser¹, Efua Ssemenda²

¹ University of Ghana, GHANA

² Mohammed V University in Rabat, MOROCCO

ABSTRACT

Ensemble learning has become a default choice for breast-cancer prediction from cytological features, but combining multiple models raises a specific concern for clinical decision support: a prediction assembled from several classifiers can be harder to explain than a prediction from one. This review synthesizes five sources to examine whether interpretability has kept pace with the field's move toward ensembles. A systematic review of 50 mammogram-classification studies finds that conventional machine learning, deep learning, and hybrid/ensemble approaches have each reached near-ceiling accuracy on their own benchmarks, and concludes that future work must balance accuracy against interpretability and resource efficiency rather than optimizing accuracy alone. Four applied studies, remarkably, all evaluate their models on the identical 569-sample Wisconsin Diagnostic Breast Cancer cytological dataset, making their results directly comparable: a Random Forest model reaches 96% accuracy with a perfect AUC of 1.00; a three-way rotating stacking ensemble of Support Vector Machine, Naïve Bayes, and K-Nearest Neighbors reaches 95% accuracy; a six-model comparison built explicitly around SHAP-based interpretability finds its Stacking Classifier reaching 98.25% accuracy with zero false positives, identifying tumor border irregularity and size as the most influential features; and a hybrid ensemble combining three classical models with three deep-learning architectures reaches 99.76% accuracy on the same data. Taken together, these four independently conducted studies on one shared benchmark show ensemble accuracy converging near a ceiling the underlying dataset may simply not allow further improvement past, while only one of the four builds interpretability into its methodology as a first-class design goal rather than an afterthought. This review argues that the field's next problem is not squeezing further accuracy from an already-saturated benchmark, but ensuring that the explanation delivered alongside an ensemble's prediction is as reliable, and as clinically actionable, as the prediction itself.

KEYWORDS: Ensemble Learning, Breast Cancer, Interpretability, Explainable AI, Clinical Decision Support, SHAP, Random Forest, Stacking Classifier

INTRODUCTION

WHY ENSEMBLES COMPLICATE THE EXPLANATION, NOT JUST THE COMPUTATION

An ensemble model combines the predictions of several individual classifiers, typically improving accuracy and robustness over any single model a well-established result this review's four applied studies each confirm in their own way. But that same combination is precisely what makes an ensemble harder to explain than a single decision tree or logistic regression: a clinician asking why a specific patient was classified as malignant is, for an ensemble, asking a question about the combined behavior of several models rather than the transparent logic of one. Explainability techniques such as SHAP were developed specifically to answer that harder question, but as this review's title suggests, applying an explanation method after the fact is a different, and not automatically equivalent, design choice from building interpretability into an ensemble's development from the start [1-7].

SCOPE: FROM CYTOLOGICAL FEATURES SPECIFICALLY, TOWARD DECISION SUPPORT BROADLY

This review's title names a specific data type, cytological features, and a specific end goal, clinical decision support, and treats the path between them as the object of study. Section 2 reviews the broader ensemble-learning landscape for breast-cancer imaging through a systematic review, establishing that near-ceiling accuracy is now common across method categories; Section 3 reviews four applied studies that, remarkably, all use the identical cytological dataset, allowing direct comparison of both their accuracy and their approach to interpretability; and Section 4 treats what these four studies collectively confirm and what they leave open about ensemble learning as clinical decision support.

REAL-WORLD GROUNDING: WHEN "INTERPRETABLE" ALSO HAS TO MEAN "REPRODUCIBLE"

A well-documented episode from outside this review's five core sources illustrates a dimension of interpretability distinct from, but closely related to, the per-prediction explanations Section 3.3 focuses on. McKinney *et al.* (2020), publishing in *Nature*, reported an AI system for breast cancer screening developed by Google Health and DeepMind that outperformed radiologists on a large international mammography dataset. Haibe-Kains, Adam, Hosny, and colleagues (2020), also in *Nature*, published a formal reply arguing that "the lack of details of the methods and algorithm code undermines its scientific value," identifying insufficient methodological transparency and unavailable implementation code as barriers that prevented the wider research

community from independently verifying or building on the reported results. This is a different failure mode from Section 3.4's undemonstrated case-level explanation: it is a failure of methodological transparency at the level of the study itself, not of explaining any individual prediction. It bears directly on this review's title, though, because Section 3's four applied studies all work from one openly available public dataset, the Wisconsin Diagnostic Breast Cancer set — a baseline transparency advantage the proprietary Google Health mammography data behind McKinney *et al.* (2020) did not share, even though this review was not able to independently confirm that all four of Section 3's studies have themselves published their model code alongside their reported results [8-13].

REVIEW SCOPE AND METHODOLOGY

This review synthesizes five sources spanning 2023 to 2026, selected on relevance and evidentiary strength rather than a fixed recency cutoff. One is a systematic review anchoring this review's broadest claims about accuracy ceilings and the interpretability-efficiency trade-off across the wider mammogram-classification literature; four are individually verified applied studies, all evaluated on the same 569-sample Wisconsin cytological dataset, which this review treats as a direct and unusually clean point of comparison rather than four unrelated results [14-17].

THE BROADER ENSEMBLE-LEARNING LANDSCAPE: A SYSTEMATIC VIEW

NEAR-CEILING ACCURACY ACROSS THREE METHOD CATEGORIES

Fajrin and Min (2025) systematically review 50 studies (2018–2025) on mammogram classification, comparing conventional machine learning, deep learning, and hybrid/ensemble approaches under PRISMA guidelines. Their most striking finding for this review's argument is not which category performs best, but that each category now reports near-ceiling accuracy on its own benchmark: an optimized Extreme Learning Machine reaches 100% accuracy on the MIAS dataset, a Vision Transformer reaches 100% accuracy on DDSM (outperforming conventional convolutional networks), and a hybrid ensemble architecture, IEUNet++, reaches 99.87% accuracy on a multi-class benchmark. Read individually, each of these looks like a solved problem; read together, as Fajrin and Min explicitly do, they suggest instead that the remaining headroom on these specific, well-worn benchmarks is nearly exhausted, and that further gains for the field will need to come from elsewhere — a conclusion directly relevant to Section 3's four studies, which push the Wisconsin dataset toward a comparable ceiling of its own.

PREPROCESSING BURDEN AS A HIDDEN COST OF THE CEILING

Fajrin and Min (2025) note that the conventional-ML and deep-learning studies reaching near-ceiling accuracy typically require extensive preprocessing, feature extraction, and hyperparameter optimization to get there — a cost that does not show

up in a headline accuracy figure but matters directly for this review's concern with clinical decision support. Section 3's four studies illustrate the same pattern on a different dataset: Sampson *et al.* (2026) apply Winsorization, square-root transformation, and feature consolidation before their models see the data at all, while Wolak *et al.* (2026) compare four separate class-imbalance-handling strategies (weighting, SMOTE, SMOTE+Tomek, undersampling) before settling on one. A clinical decision support tool is not just its final model; it is that model plus the preprocessing pipeline feeding it, and the review's finding that hybrid methods can reduce this hidden preprocessing burden is as relevant to real-world deployability as any accuracy figure.

THE REVIEW'S OWN CONCLUSION: INTERPRETABILITY AND EFFICIENCY, NOT JUST ACCURACY

Fajrin and Min (2025) draw an explicit methodological conclusion from this ceiling effect: that traditional ML and deep-learning approaches typically demand extensive preprocessing, feature extraction, and optimization to reach their reported accuracy, while hybrid methods can deliver comparable or superior performance with fewer handcrafted steps, particularly for the harder multi-class case. Their stated conclusion — that future systems require balancing model accuracy, interpretability, and resource efficiency, rather than accuracy alone — is not a generic caveat added to a methods paper; it is the review's own stated takeaway from surveying 50 studies, and it anticipates almost exactly the comparison this review's Section 3 draws out across four studies on one shared dataset.

FOUR STUDIES, ONE DATASET: A NATURAL COMPARISON

A BASELINE RESULT: RANDOM FOREST ON CYTOLOGICAL FEATURES

Siddiqui, Rahman, Kabir, Mahmud, Rashid, and Shammah (2023), the source this review's author identified directly for this topic, compare Random Forest and Gaussian Naive Bayes on 30 morphometric features from the Wisconsin Diagnostic Breast Cancer dataset (569 samples: 357 benign, 212 malignant). Random Forest — itself an ensemble of decision trees — reaches 96% accuracy with balanced precision and recall and a perfect AUC of 1.00, ahead of Gaussian Naive Bayes's 93% accuracy and AUC of 0.99. As a bagging-style ensemble evaluated with no dedicated post-hoc explainability technique applied to its own output, this study establishes both this review's accuracy baseline and, by omission, exactly the interpretability gap Sections 3.2 through 3.4 address with increasing explicitness.

A THREE-WAY ROTATING STACKING COMPARISON

Sampson, James, and Tripathi (2026) build three distinct stacking-ensemble configurations on the same Wisconsin dataset (with nine retained features after

removing fractal dimension and consolidating multicollinear variables), rotating Support Vector Machine, Naïve Bayes, and K-Nearest Neighbors between the base-learner and meta-learner roles across three model variants. Their best configuration, combining Naïve Bayes and KNN as base learners with SVM as the meta-learner, reaches 95% accuracy, 90% recall, $F1 = 93\%$, and a mean AUC of 0.97 across 10-fold cross-validation — a slightly lower accuracy than Section 3.1's Random Forest baseline, illustrating directly that stacking multiple simple models does not automatically outperform a single well-tuned ensemble method, even on identical data.

SIX CLASSIFIERS AND SHAP: INTERPRETABILITY AS A FIRST-CLASS DESIGN GOAL

Wolak, Plichta, and Orlicki (2026) provide this review's clearest example of an ensemble study built around interpretability from the outset rather than as an afterthought. Comparing six classifiers (KNN, SVM, Random Forest, CatBoost, a soft-voting classifier, and a stacking classifier with logistic-regression meta-learning) on the same Wisconsin dataset under stratified 10-fold cross-validation, their Stacking Classifier reaches the strongest overall result: 98.25% accuracy, $F1 = 0.9808$, $AUC = 0.9974$, and — notably — zero false positives, meaning no benign case was misclassified as malignant. Critically, their statistical testing (a DeLong test with Holm correction) found the AUC differences between the top-performing models were not statistically significant, a finding this review considers important: on this dataset, several different ensemble architectures are performing indistinguishably well, so the choice between them may reasonably be settled by interpretability rather than by accuracy alone. Their SHAP analysis, applied at both global and case-level detail, identifies `concave_points_worst`, `area_worst`, and `radius_worst` — measures of tumor border irregularity and size — as the most influential features, giving a clinician a concrete, feature-level account of what is driving any individual prediction.

A HYBRID ML-PLUS-DEEP-LEARNING ENSEMBLE PUSHES THE SAME BENCHMARK FURTHER

Iftikhar, Hashem, Qureshi, Rodrigues, Ali, Gonzales Medina, and López-Gonzales (2026) combine three classical machine-learning models (Logistic Regression, SVM, Random Forest) with three deep-learning architectures (AlexNet, GRU, and a bidirectional GRU) through a validation-weighted averaging mechanism, again on the same Wisconsin dataset, testing three train-test split ratios. Their best result, at a 90/10 split, reaches 99.76% accuracy with $F1 = 0.9873$, sensitivity 0.9814, and specificity 0.9913 — the highest accuracy among this review's four Wisconsin-dataset studies, and closer still to Section 2.1's observed ceiling than Section 3.3's already-strong result. The authors describe their weighted-averaging mechanism as balancing interpretability and

deep representational capacity, but this review was not able to independently verify, from the accessible content, a dedicated case-level explanation output comparable to Section 3.3's SHAP analysis; the paper's interpretability claim is architectural rather than demonstrated with the same specificity.

A SHARED BENCHMARK'S ADVANTAGES, AND ONE METHODOLOGICAL DIFFERENCE WORTH FLAGGING

Sections 3.1 through 3.4 all draw on the same 569 patient samples, but they do not all use the same features: Siddiqui *et al.* (2023) and Wolak *et al.* (2026) both retain the full set of 30 morphometric features, while Sampson *et al.* (2026) deliberately reduce to nine features after removing fractal dimension and consolidating multicollinear variables, and Iftikhar *et al.* (2026) works from the standard 30-feature version with min-max normalization applied. This means the four studies' accuracy figures, while genuinely comparable in the sense that they share the same underlying patients and diagnostic labels, are not run on strictly identical input representations — a review treating them as a perfectly controlled experiment would overstate how directly comparable they are. That Section 3.2's reduced-feature model performs only marginally below the full-feature models, rather than substantially worse, is itself a modest but genuine finding about how much redundant information the full 30-feature set may carry, consistent with Section 3.3's own finding that just three features (concave_points_worst, area_worst, radius_worst) account for most of its model's predictive signal.

FROM A PUBLIC BENCHMARK TO A DEPLOYED SCREENING TOOL: WHAT WOULD STILL BE MISSING

Section 1.3's grounding case is a useful check on how far Section 3's strong results actually travel toward deployment. McKinney *et al.*'s (2020) system was evaluated on large, real, multi-site mammography data before publication, and still drew a formal published objection over insufficient methodological transparency; none of Section 3's four studies has yet been tested at that scale, on that kind of real clinical imaging pipeline, or subjected to that level of external scrutiny. This is not a criticism specific to any one of the four — it is simply where they currently sit on the path from a public, well-understood benchmark to a clinically deployed tool: further along than a proof-of-concept, given four independent groups converging on similarly strong results, but well short of the scale, scrutiny, and reproducibility standard Section 1.3's grounding case shows even a large, well-resourced, high-profile study can fail to fully meet on its first attempt.

COMPARATIVE SYNTHESIS

Table 1. All five sources reviewed

The four applied-study rows below Fajrin and Min (2025) are not four unrelated studies — they are four independently conducted evaluations of the same 569-sample dataset, which is precisely what makes this review's comparison unusually direct rather than a synthesis across incompatible benchmarks, as is more typical of a review like this one. The final row, marked with an asterisk, is not a peer-reviewed ensemble-learning study at all but the real-world grounding case discussed in Section 1.3.

Study	Domain / dataset	Approach	Key reported result
Fajrin & Min (2025)	Mammogram classification — systematic review of 50 studies (2018–2025)	PRISMA-guided review comparing conventional ML, deep learning, and hybrid/ensemble approaches	Optimized ELM: 100% accuracy (MIAS); ViT: 100% accuracy (DDSM); hybrid IEUNet++: 99.87% (multi-class); calls for balancing accuracy, interpretability, and resource efficiency
Siddiqui <i>et al.</i> (2023)	Breast cancer diagnosis — Wisconsin cytological dataset, 569 samples	Random Forest vs. Gaussian Naive Bayes on 30 morphometric features	Random Forest: 96% accuracy, balanced precision/recall, AUC = 1.00
Sampson <i>et al.</i> (2026)	Breast cancer diagnosis — same Wisconsin dataset, 569 samples, 9 retained features	Three stacking-ensemble configurations (SVM, Naïve Bayes, KNN as base/meta learners in rotation)	Best configuration (NB+KNN base, SVM meta): 95% accuracy, 90% recall, F1 = 93%, mean AUC = 0.97
Wolak <i>et al.</i> (2026)	Breast cancer diagnosis — same Wisconsin dataset, 569 samples	Six classifiers compared (KNN, SVM, Random Forest, CatBoost, Voting, Stacking); SHAP for global and local interpretability	Stacking Classifier: 98.25% accuracy, F1 = 0.9808, AUC = 0.9974, zero false positives; SHAP ranks concave_points_worst, area_worst, radius_worst as most influential
Iftikhar <i>et al.</i> (2026)	Breast cancer diagnosis — same	Hybrid ensemble combining 3 classical	Best scenario (90/10 split): 99.76% accuracy, F1 = 0.9873,

Study	Domain / dataset	Approach	Key reported result
	Wisconsin dataset, 569 samples	ML models (LR, SVM, RF) with 3 deep learning models (AlexNet, GRU, BiGRU) via weighted averaging	sensitivity 0.9814, specificity 0.9913
McKinney <i>et al.</i> (2020) / Haibe-Kains <i>et al.</i> (2020)*	Real-world grounding case — AI breast cancer screening, proprietary mammography data	Landmark AI screening study (Nature, 2020), followed by a formal published reply on transparency and reproducibility	Reply states "the lack of details of the methods and algorithm code undermines its scientific value"; cited for real-world grounding only (Section 1.3)

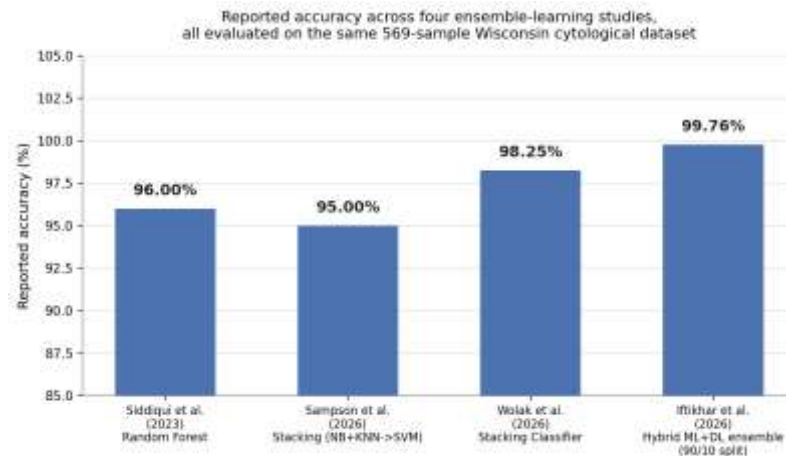


Figure 1. Reported accuracy across the four applied studies in this review, all evaluated on the identical 569-sample Wisconsin Diagnostic Breast Cancer dataset.

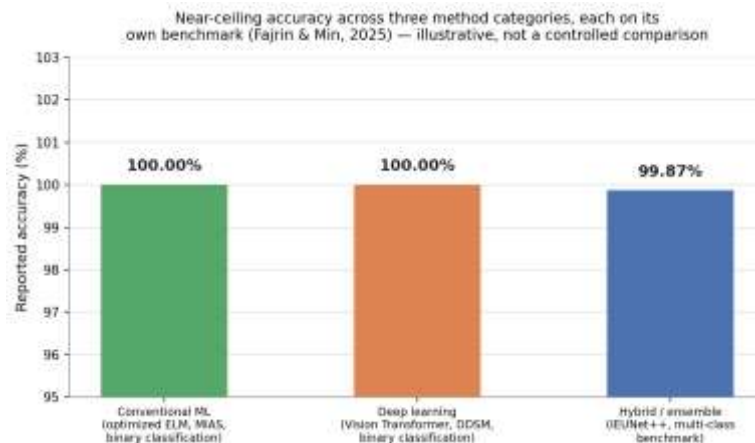


Figure 2. Best-in-class accuracy per method category from Fajrin and Min's (2025) systematic review — three different datasets/tasks, shown together only to illustrate the review's own near-ceiling finding, not as a controlled comparison (Section 2.1).

DISCUSSION

THE DATASET MAY BE CLOSER TO SOLVED THAN THE CLINICAL PROBLEM IS

Section 3's four studies report accuracy figures of 95%, 96%, 98.25%, and 99.76% on the identical Wisconsin dataset, and Section 3.3's own statistical test finds the top models' AUC differences not statistically significant — converging evidence, from two independent angles, that this specific 569-sample benchmark is approaching the same kind of ceiling Section 2.1's broader systematic review documents across mammogram datasets. This is a genuinely different claim from saying breast-cancer diagnosis from cytological features is a solved clinical problem: the Wisconsin dataset is nearly three decades old, drawn from one source, and of modest size by contemporary standards, so a near-ceiling result on it demonstrates that ensemble methods have matured against this particular benchmark, not that they have been shown to generalize to the full diversity of real-world cytological samples a deployed clinical system would actually encounter.

INTERPRETABILITY WAS OPTIONAL IN THREE OF THESE FOUR STUDIES

Only Section 3.3 (Wolak *et al.*, 2026) builds a dedicated, feature-level explanation method into its methodology and reports what it found. Section 3.1's Random Forest, though itself interpretable in principle through feature-importance scores, is not accompanied by any explanation analysis in the study reviewed; Section 3.2's stacking comparison reports accuracy and cross-validation statistics but no explanation of which features drove its predictions; and Section 3.4's hybrid ensemble describes its weighted-averaging mechanism as balancing interpretability with representational capacity but does not demonstrate a case-level explanation output this review could verify. If

interpretability is genuinely a precondition for clinical decision support, as this review's title implies, three of the four highest-performing studies on this benchmark have not yet demonstrated it to the same standard as the fourth.

WHAT SECTION 3.3'S SPECIFIC FINDINGS ADD FOR A CLINICIAN

Section 3.3's SHAP analysis is worth returning to for what it adds beyond a bare accuracy figure: identifying `concave_points_worst`, `area_worst`, and `radius_worst` as the most influential features translates a black-box prediction into a claim a clinician can check against their own reading of the cytology sample — does this case in fact show more irregular, less rounded cell-nuclei borders and larger measured area, the same features driving the model's malignant classification? Section 3.3's additional finding of zero false positives is clinically significant in its own right, since it means the study's Stacking Classifier never triggered an unnecessary biopsy or intervention for a benign case, though this review notes that a small held-out test set on a 569-sample dataset makes any single zero-false-positive result somewhat sensitive to which cases fell into that particular test split.

REGULATORY AND ETHICAL CONTEXT

The FDA's Good Machine Learning Practice principles for AI/ML-based medical devices explicitly call for interpretability appropriate to a model's intended use and users, a standard Section 5.2's finding suggests only one of this review's four applied studies has directly addressed. None of the four discusses regulatory pathways for a cytological ensemble classifier specifically, which is a reasonable scope limitation for methods-focused research, but it leaves open exactly how a model like Section 3.4's hybrid ensemble, whose interpretability claim is architectural rather than demonstrated, would satisfy that standard in practice. An ethical dimension follows directly from Section 5.1: validating four ensemble methods against the same aging, single-source benchmark risks mistaking convergence on that dataset for readiness across the more demographically and technically varied populations a real deployment would serve.

FUTURE RESEARCH DIRECTIONS

Three directions follow from the gaps this review has identified. First, given Section 5.1's finding that the Wisconsin dataset itself may be near-saturated, future ensemble-learning research on cytological features would benefit more from validation on newer, larger, and more diverse cytology datasets than from further incremental accuracy gains on this specific benchmark. Second, Section 5.2's finding that only one of four studies demonstrates case-level explainability suggests that reporting a dedicated interpretability analysis, following Section 3.3's example, should become a standard expectation for ensemble studies aimed at clinical decision support, not an optional addition. Third, Section 3.4's undemonstrated interpretability claim for its hybrid ML-

deep-learning architecture points to a specific, checkable gap: future hybrid-ensemble studies combining classical and deep models should report a feature-level or case-level explanation method with the same rigor Section 3.3 applies, rather than describing interpretability as an architectural property without demonstrating it.

THE COMPUTATIONAL COST OF THE HIGHEST-ACCURACY RESULT

Section 3.4's best-performing study also carries the highest architectural complexity among Section 3's four studies: six constituent models (three classical, three deep-learning) combined through weighted averaging, evaluated across three separate train-test split scenarios with 10-fold cross-validation. Section 3.1's Random Forest baseline, by contrast, is a single ensemble method requiring far less computational overhead to train, tune, and eventually deploy, for a difference of roughly 3.76 percentage points in accuracy (96% versus 99.76%). Whether that accuracy gain justifies six times the constituent models and the additional deep-learning infrastructure they require is a genuine deployment trade-off this review's sources do not evaluate directly, but it is the same trade-off Section 2.2 identifies at the level of the broader mammogram literature: a more complex hybrid architecture can buy real accuracy, at a real, usually unquantified, cost in complexity and resource requirements.

TRANSPARENCY AND PER-PREDICTION EXPLANATION ARE TWO SEPARATE OBLIGATIONS

Section 1.3's grounding case and Section 3.3's SHAP analysis address two different obligations that this review's title, by naming both "interpretable" and "clinical decision support," implicitly asks a deployed system to meet together. Methodological transparency, the concern Haibe-Kains *et al.* (2020) raise, is about whether other researchers and regulators can audit how a model was built and validated at all; per-prediction explainability, Section 3.3's contribution, is about whether a treating clinician can understand any one specific output. A study could in principle satisfy either without the other — publishing full code and data while still returning unexplained predictions to clinicians, or providing rich SHAP explanations for a model whose training process itself was never made auditable. None of Section 3's four studies is reported here as satisfying both simultaneously to a fully verifiable standard, which this review treats as a genuine, currently open gap in the literature rather than a minor omission.

LIMITATIONS OF THIS REVIEW

This review synthesizes five sources identified through targeted, non-exhaustive searching rather than its own registered systematic-review protocol, though it draws substantially on one larger systematic review for its broadest claims. All four applied studies in Section 3 were independently verified directly from their published results;

none required the disclosure caveats this reviewer has applied to sources in companion reviews on other topics. This review's comparison across Section 3's four studies is unusually direct because all four share one dataset, but that same shared dataset is also this review's most significant limitation: none of its conclusions about accuracy ceilings or interpretability gaps have been checked against a second, independent cytological dataset.

TOWARD ENSEMBLE LEARNING THAT EXPLAINS ITSELF BY DEFAULT

Pulling Sections 1 through 4 together, this review's title names a path from cytological features to clinical decision support, and the evidence gathered here suggests that path currently has a gap in its middle: strong, converging accuracy (Section 5.1) connects cytological features to a prediction, but only one of four studies reviewed connects that prediction back to an explanation a clinician could independently check (Section 5.2). A genuinely decision-support-ready ensemble would report Section 3.3's kind of case-level, feature-attributed explanation as a standard output alongside every prediction, not as one study's distinguishing contribution among four. Until that becomes the norm rather than the exception on this literature's own shared benchmark, ensemble learning for breast-cancer prediction will have solved the prediction half of this review's title more thoroughly than the decision-support half.

WHAT A CLINICIAN READING THIS LITERATURE TODAY WOULD ACTUALLY BE TOLD

It is worth stating plainly what a clinician or hospital administrator evaluating these four studies today would and would not learn from their published results. They would learn that several different ensemble architectures reach 95–99.76% accuracy on a well-known cytological benchmark, and, from Section 3.3 specifically, that tumor border irregularity and size are the features most consistently driving a malignant classification. They would not learn, from any single one of these four papers, how the model would perform on their own patient population, imaging equipment, or laboratory's cytology preparation; how confident to be in an individual prediction near the decision boundary rather than a clearly benign or clearly malignant case; or, for three of the four studies, why the model reached its conclusion for any specific patient at all. This is not a failure unique to breast-cancer ensemble research — Section 1.3's grounding case shows an even larger, more heavily resourced study facing comparable scrutiny — but it is the specific, concrete distance this review's evidence shows still separates "prediction" from "clinical decision support" in its title today.

CONCLUSION

Ensemble learning for breast-cancer prediction from cytological features has reached a point where four independently conducted studies, all on the same 569-sample

benchmark, report accuracy between 95% and 99.76%, with statistical testing in one of them finding the leading models' performance indistinguishable — strong evidence that this specific dataset is approaching a ceiling, echoing a broader systematic review's finding of near-ceiling accuracy across mammogram-classification method categories as well. What this literature has not yet matched is interpretability: only one of the four applied studies reviewed here builds a dedicated, feature-level explanation method into its evaluation, identifying tumor border irregularity and size as the features driving its predictions, while the other three report strong accuracy without an equivalent account of why. This review's real-world grounding case reinforces why that gap matters beyond this one benchmark: even a large, well-resourced, high-profile AI screening study can draw a formal published objection over insufficient transparency, so a smaller academic study reporting strong accuracy on a public dataset should expect at least as much scrutiny over how, and how clearly, it explains its own predictions.

REFERENCES

- [1] Fajrin, H. R., & Min, S. D. (2025). From machine learning to ensemble approaches: A systematic review of mammogram classification methods. *Diagnostics*, 15(22), Article 2829.
- [2] Karim, N. B., Tolosana, D., & Simko, R. (2025b). Responsible and Legally Compliant Enterprise Artificial Intelligence in Small-Business Lending: Governance, Explainability, and Fair-Lending Safeguards for Early-Intervention Recommendation Engines. *International Journal of Modern Computing*, 8(1), 55-66.
- [3] Shahiduzzaman, M., Sourav, M. S. A., Kabtia, M., Shammah, R. S., Islam, M. A., Chakraborty, P., ... & Adnan, M. (2026). Attention-guided quantum-inspired extreme learning machine with grey wolf optimization for rare-attack detection in IIoT networks. *Discover Artificial Intelligence*.
- [4] Haibe-Kains, B., Adam, G. A., Hosny, A., *et al.* (2020). Transparency and reproducibility in artificial intelligence. *Nature*, 586(7829), E14–E16.
- [5] Potharaju, B., & Ravipudi, J. (2025). Automated Data Validation and Resilient Infrastructure Design for Enterprise Regulatory Reporting Platforms. *International Journal of Computer Science and Engineering Innovations*, 1(1), 46-64.
- [6] Rashid, S. U., & Sattiraju, S. R. (2025a). Data-Driven Early-Warning Indicators of Small-Business Financial Distress: Integrating Operational, Transactional, and Credit Data for Decision Support in SMEs. *The Metascience*, 3(4), 53-65.
- [7] Ahmed, K. R., Sufian, M. A., Faruq, O., Kowser, S., Shammah, R. S., & Ahmmmed, M. R. (2025, July). Impact of IT product innovation on competitive advantage in the digital economy. In *2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN)* (pp. 1-6). IEEE.

- [8] Iftikhar, H., Hashem, A. F., Qureshi, M., Rodrigues, P. C., Ali, S. O., Gonzales Medina, R. I., & López-Gonzales, J. L. (2026). An intelligent hybrid ensemble model for early detection of breast cancer in multidisciplinary healthcare systems. *Diagnostics*, 16(3), Article 377.
- [9] Rahman, M. S., Islam, C. S., Mudryi, M., & Cakmak, S. (2025). Concept-Drift-Resilient Ensemble Learning for Financial Transaction Fraud Detection Under Evolving Criminal Typologies. *International Journal of Acta Informatica*, 4(1), 52-67.
- [10] McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., *et al.* (2020). International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788), 89–94.
- [11] Sampson, A., James, A., & Tripathi, V. (2026). Optimizing breast cancer prediction through stacking ensemble machine learning models: A comparative analysis. *Journal of Electrical Systems and Information Technology*, 13(1), Article 13.
- [12] Ahmed, K. R., Ahmmed, M. R., Kowser, S., Faruq, O., Shammah, R. S., & Sufian, M. A. (2025, July). The Future of AI-Powered SaaS: Market Trends and Investment Opportunities. In *2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN)* (pp. 1-6). IEEE.
- [13] Wolak, W., Plichta, A., & Orlicki, H. (2026). Interpretable ensemble learning for tumor-type prediction with a SHAP-based evaluation of CatBoost and voting classifiers. *Scientific Reports*, 16, Article 1401.
- [14] Potharaju, B., & Ravipudi, J. (2023). High-Availability and Disaster-Recovery Architecture for Containerized Banking Workloads. *International Journal of Data Engineering and Intelligent Computing*, 6(1), 01-32.
- [15] Rashid, S. U., & Sattiraju, S. R. (2025b). Multi-Institution Validation of Drift-Resilient Early-Warning Models for SME Loan Distress and Transaction Fraud: Evidence from US Community Financial Institution Pilots. *International Journal of Modern Computing*, 8(1), 40-54.
- [16] Rahman, M. S., Islam, C. S., & Mudryi, M. (2025). Interoperability Standards for Real-Time Fraud-Signal Exchange Across US Financial Institutions: Design Principles and a Reference Implementation. *The Computertech*, 47-60.
- [17] Karim, N. B., & Kovac, T. (2025). Cloud-Based Big-Data Architectures for AI-Driven Financial Planning and Analysis and Early Risk Signaling in Resource-Constrained Community Financial Institutions. *The Metascience*, 3(3), 29-40.