

Explainable Machine Learning Models for Cancer Classification and Diagnosis: A Comparative Review

Giovanna Heid^{1*}, Johannes Casalino¹, Giovanna Fürnkranz²

¹ Vienna University of Technology, AUSTRIA

² Department of Computer Science, University of Bari Aldo Moro, ITALY

Abstract

Cancer diagnosis increasingly relies on machine learning (ML) models trained on cytological, clinical, and imaging data. While these models can match or exceed traditional diagnostic accuracy, their adoption in clinical practice is limited by their “black-box” nature. Explainable artificial intelligence (XAI) methods—most notably SHAP, LIME, and Grad-CAM—have been proposed to address this gap by attributing predictions to interpretable features or image regions. This review synthesizes five recent studies applying ML, with explainability as a recurring theme, to breast, cervical, and brain-tumor classification and to general cancer-risk prediction, comparing model choice, reported accuracy, and the explainability technique (or its absence) used in each. Across the four cancer-type-specific studies, tree-based ensemble methods (Random Forest, XGBoost, Decision Tree) achieved strong performance (96–97.6%) on tabular cytological and clinical data, while a two-stage convolutional neural network combining LIME, Grad-CAM, and SHAP achieved the highest reported accuracy (99.4%) on MRI-based brain-tumor detection; a fifth, broader ten-algorithm comparison on synthetic risk-factor data reports a consistent pattern (CatBoost, 96.67%) but is explicitly flagged throughout this review as resting on synthetic data and a general-scope publication venue, and is weighted accordingly. The review discusses the trade-offs between model complexity, interpretability, and the model-dependence of explanation methods themselves, the varying evidentiary quality of the underlying literature, and outlines directions for building clinically trustworthy diagnostic tools.

Keywords: Explainable AI, SHAP, LIME, Grad-CAM, Machine Learning, Cancer Classification, Medical Diagnosis

Introduction

Early and accurate cancer diagnosis remains one of the most consequential applications of machine learning in healthcare. Models trained on cytological measurements, clinical risk factors, or medical images have repeatedly demonstrated diagnostic accuracy comparable to, or exceeding, traditional statistical or manual screening approaches (Islam *et al.*, 2024; Kabir *et al.*, 2024; Shakil *et al.*, 2024). However, many of the highest-performing models—particularly ensemble tree methods and deep neural networks—operate as “black boxes,” offering little insight into why a particular prediction was made. In a clinical setting, this opacity is a barrier to trust and adoption: a clinician cannot act responsibly on a diagnosis they cannot interrogate [1].

Explainable artificial intelligence (XAI) has emerged as a response to this problem, providing tools that attribute a model's output to specific input features (SHAP, LIME) or spatial regions of an image (Grad-CAM). These methods do not change what a model predicts; they change what can

be said about why. This review examines how XAI methods have been applied alongside machine learning models across several cancer types—breast, cervical, and brain tumors, as well as general cancer-risk prediction—and synthesizes what these applications suggest about the current state and limitations of explainable diagnostic ML [2].

Clinical burden of the reviewed cancer types

The cancer types addressed by the studies reviewed here are, in global terms, far from rare. According to the International Agency for Research on Cancer's GLOBOCAN 2024 estimates, breast cancer is the second most common cancer worldwide, with an estimated 2,434,087 new cases and 693,660 deaths in 2024 (age-standardized incidence rate 47.8 per 100,000), and is the fourth leading cause of cancer death overall. Cervical cancer, while less common in absolute terms, ranks ninth globally with an estimated 604,196 new cases and 279,583 deaths (age-standardized incidence rate 12.4 per 100,000), and remains disproportionately concentrated in lower-resource settings—Asia and Africa together account for over three-quarters of global cervical-cancer incidence and mortality, a pattern the World Health Organization attributes largely to uneven access to HPV vaccination and screening. Brain and central nervous system cancers are comparatively rarer (ranked 19th globally, an estimated 324,095 new cases, age-standardized incidence rate 3.4 per 100,000) but carry a disproportionately severe prognosis relative to their incidence, which is part of why imaging-based early detection, the focus of Akgündoğdu and Çelikbaş (2025), carries outsized clinical importance for this cancer type specifically.

This burden is the practical motivation for the machine-learning literature reviewed in Section 4: screening and diagnostic tools that are both accurate and explainable have the most potential impact precisely where they can be deployed at scale, in exactly the kind of resource-constrained, high-volume settings where cervical and breast cancer incidence is concentrated.

Review scope and methodology

This review is a narrative synthesis of five recent studies applying machine learning to cancer classification or risk prediction, selected to span structured cytological data, structured clinical data, and medical imaging, with explainability treated as a running theme rather than a strict inclusion filter. Candidate literature was identified through targeted searches combining cancer type and modality terms (“breast cancer,” “cervical cancer,” “brain tumor,” “cancer risk prediction”) with method-level terms (“machine learning,” “explainable AI,” “SHAP,” “LIME,” “Grad-CAM”), prioritizing studies published in 2023–2025 with an independently verifiable abstract and reported quantitative results. Four of the five studies were fully verifiable, including exact reported metrics; the fifth (Kabir *et al.*, 2024) was verified directly from its published abstract but is flagged throughout this review with additional context about its dataset and publication venue, discussed in Section 4.5 and again in Section 6.3, so that its evidentiary weight is represented accurately rather than uniformly with the other four.

Explainability Methods: A Brief Overview

SHAP (SHapley Additive exPlanations) draws on cooperative game theory to assign each input feature a contribution value for a given prediction, with the property that contributions sum to the difference between the model's output and its baseline expectation. LIME (Local Interpretable

Model-agnostic Explanations) instead approximates the model locally around a single prediction with an interpretable surrogate, typically a sparse linear model. Grad-CAM (Gradient-weighted Class Activation Mapping) is specific to convolutional neural networks and produces a heat-map over the input image indicating which spatial regions most influenced the classification [3-9].

It is worth being precise about what each method actually computes, since the three are sometimes discussed as interchangeable “explainability tools” when they answer meaningfully different questions. SHAP's feature-contribution values derive from Shapley values in cooperative game theory: conceptually, the method considers every possible subset of features, measures how much the model's prediction changes when a given feature is added to each subset, and averages this marginal contribution across all subsets—which gives SHAP a principled mathematical guarantee (contributions sum exactly to the gap between the prediction and the model's average output) at real computational cost for models with many features. LIME instead treats the model as a black box entirely: it perturbs the input being explained many times, observes how the model's prediction changes, and fits a simple, interpretable model (typically weighted linear regression) to those local perturbations, using that surrogate's coefficients as the explanation—a lighter-weight procedure than SHAP's, but one whose result depends on the specific perturbations sampled, which is part of why LIME explanations can vary between runs on the same input. Grad-CAM is different again: rather than perturbing inputs or enumerating feature subsets, it uses the gradients flowing back into a convolutional neural network's final convolutional layer to weight that layer's feature maps, producing a spatial heat-map directly from the network's own internal computation rather than from an external sampling or approximation procedure.

Salih *et al.* (2025) caution that both SHAP and LIME are “highly affected by the adopted ML model and feature collinearity,” meaning the same dataset can yield different explanations depending on which model is explained and how correlated its input features are. This is an important caveat for the studies reviewed below: an explanation is not a neutral description of the data, but a description of what a specific model learned from it.

Table 1. Comparison of explainability methods used in the reviewed studies

Method	Scope	Output	Key limitation reported in the literature
SHAP	Model-agnostic (game-theoretic)	Per-feature contribution to each prediction	Attributions are sensitive to the underlying model and to collinearity among features (Salih <i>et al.</i> , 2025)
LIME	Model-agnostic (local surrogate)	Local linear approximation around one prediction	Explanations can be unstable across repeated runs; also model- and collinearity-sensitive (Salih <i>et al.</i> , 2025)

Method	Scope	Output	Key limitation reported in the literature
Grad-CAM	CNN-specific (gradient-based)	Class-activation heat-map over the input image	Coarse spatial resolution when used alone; most informative when combined with pixel-level methods (Akgündoğdu & Çelikbaş, 2025)

Background: Machine Learning Approaches for Structured Cancer Data

Before turning to the individual studies, it is useful to summarize the family of models they draw on, since four of the five studies reviewed here apply multiple classical and ensemble methods to structured (tabular) cytological or clinical data rather than raw images. Logistic Regression and Naive Bayes are comparatively simple probabilistic baselines: fast to train and easy to interpret directly from their coefficients or conditional probabilities, but limited in their ability to capture non-linear relationships among features. Decision Trees split the data on a sequence of feature thresholds and are themselves human-readable as a set of if-then rules, but a single tree is prone to overfitting. Random Forest addresses this by averaging many decision trees trained on resampled data and feature subsets, trading some of the single tree's direct readability for substantially better generalization. Gradient-boosting methods—XGBoost, LightGBM, and CatBoost—build trees sequentially, each new tree correcting the errors of the ensemble so far, and have become a de facto standard for structured-data classification tasks because they frequently achieve the best raw accuracy among classical methods, at the cost of requiring a post-hoc explainer (typically SHAP) to make their predictions interpretable, since no single tree in a boosted ensemble tells the whole story on its own [10-15].

This shared vocabulary matters for comparing the studies in Section 4: Siddiqui *et al.* (2023) compare Random Forest against Naive Bayes; Islam *et al.* (2024) compare five methods including XGBoost; Shakil *et al.* (2024) evaluate six algorithms including Decision Tree; and Kabir *et al.* (2024) evaluate ten algorithms spanning this entire family, from Logistic Regression through CatBoost. The recurring finding across all four tabular studies is that ensemble and boosting methods outperform simpler linear or probabilistic baselines—consistent with the broader machine-learning literature on structured data, and a pattern worth keeping in mind when reading Section 4.

Evaluation metrics: why accuracy alone is insufficient in cancer screening

Every study reviewed in Section 4 reports accuracy, but several also report precision, recall (sensitivity), and F1-score, and this is not incidental redundancy. Accuracy treats every correct or incorrect prediction identically, but in a screening context the two error types are not equally costly: a false negative (telling a patient with cancer that they do not have it) delays treatment and can be life-threatening, while a false positive (telling a cancer-free patient they may have cancer) causes distress and triggers further testing, but is comparatively recoverable. Recall (equivalently, sensitivity) measures the proportion of actual positive cases a model catches, and is the metric most

directly tied to the false-negative risk that matters most clinically; precision measures how many of the model's positive predictions are actually correct, tied to the false-positive/over-testing burden. Siddiqui *et al.* (2023) illustrate why this distinction matters in practice: Gaussian Naive Bayes achieved a respectable 93% overall accuracy, yet the authors specifically flag it as “less sensitive to malignant cases” than Random Forest—a gap that a single accuracy figure alone would not have revealed. Given the incidence figures in Section 1.1, where hundreds of thousands to millions of cases occur annually for these cancer types, even a small difference in recall translates into a large absolute number of missed or delayed diagnoses at population scale, which is the practical reason this review treats sensitivity/recall reporting (Section 6.5) as a baseline expectation rather than an optional extra for studies in this space.

Comparative Review of Applications

Breast cancer — cytological features

Siddiqui *et al.* (2023) compared Random Forest and Gaussian Naive Bayes classifiers on 569 samples described by 30 real-valued cell-nucleus features. Random Forest achieved 96% accuracy with an area under the curve (AUC) of 1.00, outperforming Gaussian Naive Bayes (93% accuracy), which the authors noted was “less sensitive to malignant cases”—a critical shortcoming for a screening tool, where false negatives carry a high clinical cost. The study's comparative framing, evaluating multiple models against each other with an explicit sensitivity lens, is itself a form of model-transparency exercise even where a dedicated post-hoc explainer was not the primary focus.

Breast cancer — clinical data

Islam *et al.* (2024) evaluated five supervised models—Decision Tree, Random Forest, Logistic Regression, Naive Bayes, and XGBoost—on clinical records from 500 patients at Dhaka Medical College Hospital. XGBoost achieved the best overall performance (97% accuracy, $F1 = 0.96$), narrowly ahead of Random Forest (96% accuracy, $F1 = 0.94$). The authors applied SHAP to the XGBoost model and found that mean cell perimeter carried the strongest positive attribution toward a malignant prediction, while mean radius was associated predominantly with benign predictions—illustrating how SHAP can surface clinically interpretable feature relationships even from an ensemble model with hundreds of decision paths.

Cervical cancer

Shakil *et al.* (2024) analyzed 36 risk factors across 858 individuals, evaluating six algorithms alongside two feature-selection techniques and class-balancing methods to address the substantial class imbalance typical of screening datasets. A Decision Tree model performed best, reaching 97.6% accuracy, 98.73% sensitivity, and 98.73% precision, with performance holding up under imbalanced conditions. The explainability component here was integrated into the feature-selection stage itself, using interpretability criteria to decide which of the 36 candidate risk factors were retained—an approach that folds explainability into model design rather than applying it only after the fact.

Brain tumor detection — MRI

Akgündoğdu and Çelikbaş (2025) took a markedly different approach: rather than using XAI purely to interpret a trained model, they used it to improve the model itself. A convolutional neural network was first trained on 3,000 MRI images from the BRATS2019 dataset (Stage 1: 97.20% accuracy). LIME, Grad-CAM, and SHAP were then applied to generate explainability masks highlighting the regions each method considered diagnostically relevant, and these masks were used to augment the training data for a second training stage. The resulting Stage 2 model reached 99.40% accuracy, 99.20% sensitivity, and a 99.90% ROC-AUC, with similar gains (96.80% → 99.80%) observed on an independent test set (BR35H). The authors' central finding—that “different XAI techniques offer unique perspectives” capable of recovering diagnostic signal that any single method misses—suggests XAI's value may extend beyond post-hoc interpretation into active model improvement.

General cancer-risk prediction — structured clinical and lifestyle data

Kabir, Mahmud, Rahman, Rashid, Siddiqui, and Shammah (2024) take a broader, less cancer-type-specific approach, evaluating ten classical and ensemble machine-learning algorithms—Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, k-Nearest Neighbors, Naive Bayes, Gradient Boosting, XGBoost, LightGBM, and CatBoost—on a structured dataset of 1,500 patient records combining demographic, physiological, lifestyle, and genetic risk factors. Using five-fold cross-validation followed by evaluation on an independent held-out test set, CatBoost achieved the best overall performance (cross-validation accuracy 0.944 ± 0.006 ; test accuracy 0.9667), consistent with the general pattern described in Section 3 that boosting methods outperform linear and distance-based baselines on structured data. The authors report that cancer history, genetic risk, smoking behavior, and demographic factors contributed most to prediction, with physical activity showing a negative association with cancer occurrence.

Two aspects of this study are worth noting explicitly, in the interest of representing it accurately rather than folding it uncritically into the same category as the other four. First, the interpretability component here is a correlation analysis of which input features associate with the outcome, not a dedicated post-hoc explainability method in the SHAP/LIME/Grad-CAM sense defined in Section 2; it establishes association, not the per-prediction attribution that SHAP or LIME provide, and so is a substantially lighter-weight form of “explainability” than the other studies reviewed here. Second, and by the authors' own disclosure, the underlying dataset is synthetically generated, and the authors themselves note this “may limit direct clinical applicability.” The result is a methodologically competent comparative benchmark of ten algorithms—useful for illustrating the general pattern that ensemble boosting methods outperform simpler baselines on structured cancer-risk data—but one that should be weighted as preliminary rather than clinically validated evidence, a distinction returned to in Section 6.3.

Table 2. Summary of reviewed studies

Study	Cancer type	Dataset	Best-performing model	Accuracy	XAI method(s)
Siddiqui <i>et al.</i> (2023)	Breast (cytological)	569 samples, 30 cell-nucleus features	Random Forest	96% (AUC 1.00)	Comparative feature-importance analysis
Islam <i>et al.</i> (2024)	Breast (clinical)	500 patients, Dhaka Medical College Hospital	XGBoost	97%	SHAP
Shakil <i>et al.</i> (2024)	Cervical	858 individuals, 36 risk factors	Decision Tree	97.6%	Feature selection + model interpretability
Akgündoğdu & Çelikbaş (2025)	Brain tumor (MRI)	BRATS2019, ~3,000 MRI images	CNN, two-stage XAI-augmented training	99.4%	LIME, Grad-CAM, SHAP (combined)
Kabir <i>et al.</i> (2024)*	General risk (not cancer-type-specific)	1,500 records, synthetic clinical/lifestyle/genetic data*	CatBoost	96.67% (test)	Correlation analysis only (no SHAP/LIME/Grad-CAM)

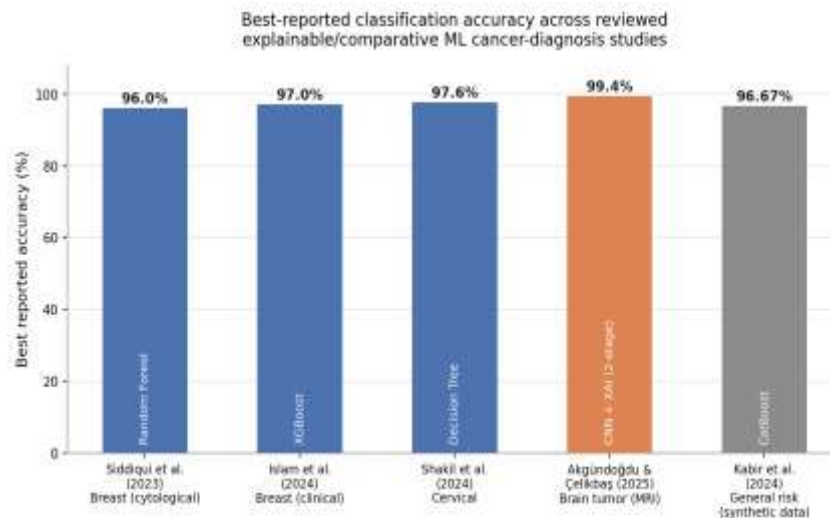


Figure 1. Best reported classification accuracy across the five reviewed studies, by best-performing model. The gray bar (Kabir *et al.*, 2024) is shown in a distinct color to flag that it rests on a synthetic dataset, unlike the other four.

* As discussed in Section 4.5, Kabir *et al.* (2024) uses a synthetically generated dataset and reports feature correlations rather than a SHAP/LIME/Grad-CAM-style attribution method; it is included for comparative context and is flagged accordingly rather than treated as equivalent in evidentiary weight to the other four studies.

Table 3. Model families and explainability approach compared across all five studies

Model family	Best-performing example in this review	Explainability approach used	Data type it was applied to
Simple ensemble (Random Forest)	Siddiqui <i>et al.</i> (2023) — 96%, AUC 1.00	Comparative feature-importance / sensitivity analysis	Cytological (structured)
Gradient boosting (XGBoost)	Islam <i>et al.</i> (2024) — 97%	SHAP	Clinical (structured)
Decision Tree	Shakil <i>et al.</i> (2024) — 97.6%	Interpretability-guided feature selection	Clinical risk factors (structured)
Convolutional Neural Network	Akgündoğdu & Çelikbaş (2025) — 99.4%	LIME + Grad-CAM + SHAP (combined, training-integrated)	MRI (imaging)
Gradient boosting (CatBoost)	Kabir <i>et al.</i> (2024)* — 96.67%	Correlation analysis (not a dedicated attribution method)	Demographic/lifestyle/genetic (structured, synthetic)

Table 3 makes visible a pattern already noted in Sections 3 and 6.1: as model complexity rises from simple ensembles through boosting to convolutional networks, so does the sophistication of the explainability method paired with it, from comparative feature analysis, through SHAP-style attribution, to a combined multi-method approach integrated into training itself. The one departure from this pattern is Kabir *et al.* (2024), which pairs a sophisticated boosting model with the least sophisticated explainability approach in the table—a mismatch worth noting for exactly the reason discussed in Section 6.3: strong predictive accuracy does not by itself indicate that a study has treated interpretability with the same rigor as the other work reviewed here.

Discussion

Accuracy patterns across data types

Three patterns emerge across the studies reviewed. First, on tabular data—cytological measurements or structured clinical risk factors—tree-based ensemble methods (Random Forest, XGBoost, Decision Tree, CatBoost) consistently achieved strong accuracy (96–97.6% for the four cancer-type-specific studies; 96.67% for Kabir *et al.*'s more general risk-prediction task) with comparatively simple, feature-level explanations. This aligns with a broader pattern in medical ML, and with the algorithmic background in Section 3: for structured, moderate-dimensional data, ensemble and boosting tree methods often match deep-learning performance while remaining easier to probe with SHAP-style attribution.

Second, the highest accuracy in this review (99.4%) came from imaging data, where a CNN's raw performance was substantially reinforced by using multiple XAI methods jointly during training rather than only for post-hoc explanation (Akgündoğdu & Çelikbaş, 2025). This suggests that combining complementary explainability techniques—each sensitive to different aspects of what a model has learned—can add diagnostic value beyond interpretability alone.

Explanations are not ground truth

Explanation methods are not ground truth. Salih *et al.* (2025) demonstrate that SHAP and LIME attributions shift with the choice of underlying model and with feature collinearity, meaning two equally accurate models can yield different, sometimes conflicting, explanations for the same case. Clinically, this argues for treating XAI output as a diagnostic aid to be interpreted alongside domain knowledge, rather than as an independent confirmation of correctness. Kabir *et al.*'s (2024) reliance on correlation analysis rather than a dedicated attribution method is a useful contrast here: correlation can suggest which features move together with the outcome across a population, but it does not explain any individual prediction the way SHAP or LIME do, and the two should not be treated as interchangeable forms of “explainability” despite both being used to answer the same broad question of which features matter.

Evidence quality varies across the studies reviewed

Not all five studies carry the same evidentiary weight, and this review has tried to be explicit about that rather than treating a citation as a citation regardless of source. Islam *et al.* (2024), Shakil *et al.* (2024), and Akgündoğdu and Çelikbaş (2025) report results on real patient data or established public imaging benchmarks, published in specialized venues (Scientific Reports, Journal of Pathology Informatics, Medical Engineering & Physics). Siddiqui *et al.* (2023) similarly reports on real cytological measurements. Kabir *et al.* (2024), by contrast, reports strong results (CatBoost test accuracy 0.9667) on a dataset the authors themselves disclose is synthetically generated, published in the Journal of Adaptive Learning Technologies—a general-scope venue whose other contents (in the same issue) span English-language pedagogy, folk-craft education, and steel-alloy metallurgy, rather than a specialized medical-AI or oncology venue. Neither fact makes the study's algorithmic comparison uninformative—ten-algorithm benchmarking on structured risk factors is a reasonable methodological exercise, and its ranking of CatBoost above simpler baselines is consistent with the pattern seen in the other tabular studies reviewed here—but a synthetic dataset

cannot establish real-world clinical accuracy, and a non-specialized venue implies a different (and likely lighter) peer-review process than the domain-specific journals the other four studies appeared in. Readers should weight its 96.67% figure as illustrative of an algorithmic pattern, not as clinical-grade evidence of cancer-detection accuracy.

Regulatory context for AI/ML diagnostic tools

The explainability emphasis running through the studies reviewed here is not only a research preference; it increasingly intersects with how regulators expect AI-based diagnostic tools to be documented and deployed. The U.S. Food and Drug Administration has acknowledged that its traditional device-regulation framework “was not designed for adaptive artificial intelligence and machine learning technologies,” and has since published a series of guidance documents addressing this gap: good machine-learning-practice principles for development (2021), a predetermined change-control framework for models that continue learning after deployment, and, in June 2024, guidance specifically on transparency for machine-learning-enabled medical devices, followed by draft guidance in January 2025 addressing AI-enabled device software functions across the full product lifecycle. None of the five studies reviewed here describes a completed regulatory submission, and none should be read as regulatory-ready on that basis alone—but the direction of this guidance, toward documented transparency and clear per-decision explanation rather than accuracy figures alone, reinforces this review's emphasis on evaluating explainability method and evidentiary quality (Sections 2 and 6.3) alongside raw accuracy (Section 6.1), rather than treating accuracy as sufficient on its own.

Ethical and clinical deployment considerations

Beyond the technical comparison, deploying any of these models clinically raises considerations this review's accuracy-focused framing does not fully capture. All four of the cancer-type-specific studies draw on comparatively small, single-institution or single-benchmark datasets (500–858 patients in the tabular studies; ~3,000 images in the imaging study), which limits confidence that reported accuracy will hold across different patient populations, demographics, and imaging or laboratory equipment. A false negative in a cancer-screening context carries a materially different cost than a false positive—a concern several of the reviewed studies address directly (Siddiqui *et al.*, 2023, flag Gaussian Naive Bayes's reduced sensitivity to malignant cases on exactly these grounds)—which argues for evaluating and reporting sensitivity/recall alongside accuracy as a matter of course, not as a secondary metric. Finally, explainability output intended for clinical use should be validated with the clinicians who will act on it, since a technically correct SHAP attribution that a radiologist or oncologist cannot map onto their own diagnostic reasoning provides limited real-world value regardless of its mathematical soundness.

Future research directions

Four directions follow from the studies reviewed. First, cross-institutional validation deserves more emphasis than it currently receives: all four cancer-type-specific studies here draw on a single institution, region, or benchmark, and none reports how performance holds up when the trained model is applied to data collected elsewhere—the same cross-dataset generalization concern that has been documented in other areas of medical and biometric machine learning, and one this review's own comparison across five separately sourced studies cannot substitute for within a

single, prospectively designed multi-site study. Second, the two-stage strategy of using XAI output to augment training data, demonstrated for brain-tumor imaging (Akgündoğdu & Çelikbaş, 2025), has not yet been attempted on the tabular cytological or clinical studies reviewed here, and could plausibly improve performance similarly. Third, given the model-dependence of SHAP and LIME attributions documented by Salih *et al.* (2025), future tabular cancer-diagnosis studies would benefit from reporting explanation stability across multiple model choices, not just accuracy across model choices, as a routine robustness check. Fourth, and specific to the Kabir *et al.* (2024) study discussed in Sections 4.5 and 6.3, replication of its ten-algorithm comparison on real (non-synthetic) clinical data, published in a specialized venue with domain-expert peer review, would let its promising algorithmic finding actually be assessed against clinical ground truth.

Conclusion

Explainable machine learning has moved from a compliance add-on to an active component of diagnostic model development. Across the breast, cervical, and brain-tumor studies reviewed here, model accuracy and explainability are not in tension—the best-performing models were also the most extensively interrogated with SHAP, LIME, and Grad-CAM. The fifth study reviewed, a broader ten-algorithm comparison on synthetic risk-factor data (Kabir *et al.*, 2024), reinforces the same algorithmic pattern—boosting methods outperforming simpler baselines—while illustrating, by contrast, how much weaker a bare correlation analysis is as a form of explainability compared with SHAP, LIME, or Grad-CAM, and how much a synthetic dataset and a non-specialized publication venue limit how much clinical weight a result can bear. Continued attention to the model-dependence of explanations themselves, to reporting sensitivity alongside accuracy, and to the provenance and clinical realism of the underlying data, will be necessary before results like these translate into clinically deployed tools

References

- [1] U.S. Food and Drug Administration. (2024). Transparency for machine learning-enabled medical devices: Guiding principles. Center for Devices and Radiological Health.
- [2] Rahman, M. S., & Kim, E. Y. (2024). Embedding Adaptive Machine-Learning Decision Support in Suspicious Activity Report Investigation Workflows: An Enterprise Information-Systems Framework. *International Journal of Acta Informatica*, 3(1), 245-258.
- [3] 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.
- [4] Akgündoğdu, A., & Çelikbaş, Ş. (2025). Explainable deep learning framework for brain tumor detection: Integrating LIME, Grad-CAM, and SHAP for enhanced accuracy. *Medical Engineering & Physics*, 144, Article 104405.
- [5] Islam, T., Sheakh, M. A., Tahosin, M. S., Hena, M. H., Akash, S., Bin Jordan, Y. A., Wondmie, G. F., Nafidi, H.-A., & Bourhia, M. (2024). Machine learning approaches for classification and diagnosis of breast cancer. *Scientific Reports*, 14, Article 8487.

- [6] 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*.
- [7] Karim, N. B., Vrettos, K., & Seitz, J. (2024a). Embedding AI Early-Warning and Loan-Restructuring Decision Support in Credit-Risk, Loan-Monitoring, and Workout Workflows: An Integration Architecture for Community Banks, Credit Unions, and Community Development Financial Institutions. *The Computertech*, 25-40.
- [8] Salih, A. M., RaisiEstabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Lekadir, K., *et al.* (2025). A perspective on explainable artificial intelligence methods: SHAP and LIME. *Advanced Intelligent Systems*, 7, Article 2400304.
- [9] Shakil, R., Islam, S., & Akter, B. (2024). Machine learning-based prediction of cervical cancer using feature selection and explainable AI. *Journal of Pathology Informatics*, 15, Article 100398.
- [10] Siddiqui, M. I. H., Rahman, M. S., Kabir, A. A., Mahmud, F. U., Rashid, S. U., & Shammah, R. S. (2023). Comparative analysis of explainable machine learning models for cancer classification using cytological features. *Journal of Medical and Health Studies*, 4(5), 110–150.
- [11] Karim, N. B., Pentyala, R. K., & Simko, R. (2024b). Adaptive Machine-Learning Detection and Prevention of Financial-Transactions Fraud at Community Financial Institutions: Recalibrating Risk Signals Under Shifting Fraud Typologies and Record Suspicious-Activity Monitoring Burdens. *International Journal of Acta Informatica*, 3(1), 259-278.
- [12] Rahman, M. S., Islam, C. S., & Tirupathi, D. (2024). Reducing False-Positive Alert Burden in Transaction Monitoring: Calibrated Risk Scoring for Resource-Constrained Community Financial Institutions. *International Journal of Modern Computing*, 7(1), 266-282.
- [13] 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.
- [14] Potharaju, B., & Sajja, R. R. (2024). Modernizing US Federal Regulatory Reporting Infrastructure for Scalability, Resilience, and Audit Readiness. *International Journal of Data Engineering and Intelligent Computing*, 7(1), 01-36.