IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 31, No. 8, August, 2026, pp. 2765-2776
Journal homepage: http://iieta.org/journals/isi

Detecting and Mitigating Data Leakage in Breast Cancer Prediction Using an Explainable Machine Learning Framework with XGBoost and SHAP

Karrar Abdullah Mohammed Habeeban[1], Aseel Faisal Alshaibanee[2], Marwah Habiban[3], Abbas Fadhil Mahdi[4*]

[1] Department of Computer Science, Faculty of Education for Women, University of Kufa, Najaf 54001, Iraq
[2] Department of Computer Science, Faculty of Computer Science and Mathematics, University of Kufa, Najaf 54001, Iraq
[3] Department of Artificial Intelligence, Faculty of Computer Science and Mathematics, University of Kufa, Najaf 54001, Iraq
[4] Department of Mathematics, Faculty of Computer Science and Mathematics, University of Kufa, Najaf 54001, Iraq

Corresponding Author Email: abbassf.wahab@uokufa.edu.iq

https://doi.org/10.18280/isi.310821

Received: 20 March 2026
Revised: 25 May 2026
Accepted: 3 June 2026
Available online: 31 August 2026

Keywords:
breast cancer, clinical machine learning, data leakage, explainable artificial intelligence, SHapley Additive exPlanations, Extreme Gradient Boosting

ABSTRACT

Machine learning models have shown considerable potential for clinical prediction; however, their reported performance can be substantially affected by data leakage, leading to unreliable estimates of predictive capability. This study proposes an explainable leakage-aware framework for identifying and mitigating data leakage in breast cancer prediction models. The proposed framework integrates data preprocessing, feature auditing, Extreme Gradient Boosting (XGBoost)-based classification, and SHapley Additive exPlanations (SHAP) analysis to examine whether highly influential predictors represent clinically available information or outcome-related leakage. Experiments were conducted using two breast cancer datasets, including the Breast Cancer Wisconsin dataset from Kaggle and a clinical dataset from Mendeley Data. The initial model trained on the Mendeley dataset achieved an accuracy of 0.9710 and receiver operating characteristic-area under the curve (ROC-AUC) of 0.9968; however, SHAP analysis revealed that several dominant predictors were associated with post-outcome information, indicating potential target leakage. After removing leakage-prone variables, the model performance decreased to a more realistic level, achieving an accuracy of 0.8950 and ROC-AUC of 0.9416. Additional experiments using random forest (RF), support vector machine (SVM), and CatBoost confirmed that leakage effects were not limited to a specific algorithm. The results demonstrate that explainability-based auditing can support more reliable evaluation of clinical machine learning models by identifying misleading predictive patterns caused by inappropriate features.

1. INTRODUCTION

Breast cancer is the most common cancer in women worldwide and is still one of the main causes of cancer death. According to the updated cancer statistics for 2020, there were more than 2.3 million new cases and approximately 685,500 deaths from breast cancer globally [1], which evidences the urgent need for powerful diagnostic and prognostic approaches that enrich clinical decision-making and personalize therapy.

Machine learning approaches have recently been applied to breast cancer in diagnosis, prognosis, and treatment-response estimation areas [2-4]. Several algorithms, including logistic regression, random forest (RF), support vector machines (SVM), and gradient boosting algorithms such as Extreme Gradient Boosting (XGBoost), achieved promising predictive results on clinical-pathological-imaging-derived datasets [2, 4, 5]. Due to their ability to capture complex, non-linear relationships in the data as well as handle heterogeneous high-dimensional features in a natural manner with an empowered framework for a fully data-driven clinical decision-making process, these approaches have become increasingly popular tools among researchers in oncology.

However, despite these advances, increasing criticism has recently been raised over the methodological quality, transparency, and clinical applicability of machine learning-based healthcare studies [6-8]. High predictive performance of a machine learning algorithm does not equal clinical utility or generalizability to real-world patient groups. However, numerous medical artificial intelligence (AI) systems were found to have reproducibility problems related to inappropriate validation procedures, methodological flaws, and hidden biases that might artificially increase predictive performance [6, 8]. This has increased the need for a framework that reflects reliable, clinically credible AI technology in health care [9].

Data leakage, which is a methodological issue [6, 10], presents one of the most serious and unrecognized threats to trustworthy predictive modelling within clinical machine learning. Data leakage is defined as the inclusion of information that should not have been available when using data that would not be available at prediction time in the final

prediction-making process, for example during training or testing of a model [10]. This is an especially problematic issue in a clinical scenario where patient records typically contain longitudinal follow-up variables, treatment outcomes or survival indicators, or post-diagnostic measurements that can act as indirect encoders for the target label [8, 11]. As a result, machine learning models may instead learn future outcome information rather than clinically meaningful prognostic patterns, ultimately leading to unrealistic predictive performance and poor generalizability in the medical domain [6, 10].

Machine learning research has identified various forms of data leakage, including target leakage, temporal leakage, and preprocessing leakage [10, 12]. Especially in clinical prediction tasks, target leakage is problematic because variables that are linked to the outcome can be included when the model is created accidentally. Temporal leakage appears when future clinical information comes into the training dataset, while preprocessing (or pipeline) leakage takes its place when average methods of scaling, normalization, imputation, or feature encoding are applied prior to suitable separation between train and test datasets [10, 12]. The most prevalent types of leakage found in clinical machine learning applications are summarized in Table 1.

Table 1. Common types of data leakage in clinical machine learning

Leakage Type	Description	Example in Clinical Data
Target leakage	Features directly related to the prediction target are included as predictors.	Overall survival status, patient vital status
Temporal leakage	Information collected after the prediction time point is used during model training.	Survival months, follow-up outcomes
Preprocessing leakage	Information from validation or test data influences preprocessing steps.	Scaling or encoding before train–test split

Explainable artificial intelligence (XAI) methods help study the transparency and interpretability of these models, which has become more important for clinical applications of healthcare machine learning [13, 14]. Of the many existing approaches, SHapley Additive exPlanations (SHAP) have become one of the most popular model-agnostic interpretive techniques for machine learning-based predictions [15]. SHAP offers global and local interpretations by estimating the contribution of each feature to model predictions. In oncology research, SHAP has mainly been applied in the context of feature importance and model interpretation [13]. However, its applications in auditing methods for implicit data leakage and unreliable predictive patterns are limited.

Newer guidelines for clinical AI, such as Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD)-AI and trustworthy AI frameworks [9, 16], highlight the importance of validation that avoids relevant leakage and transparent reporting. These approaches highlight the demand for methodological rigor, transparency of analytic techniques, and reproducibility, as well as clinically relevant assessment methods that are crucial to enable and allow efficient (and safe) translation of machine learning systems toward real-world health care models. Yet in that field, while the knowledge about reproducibility and reliability of AI output is increasing, there have been no definite studies of data leakage as a principal methodological focus in breast cancer prediction. Most existing studies focus on optimizing predictive performance and are not systematically validated to determine whether the learned patterns are clinically meaningful or temporally appropriate.

Motivated by these challenges, this work examines where data leakage occurs and how to detect and mitigate it with a specific focus on clinical machine learning models for breast cancer prediction. Unlike traditional predictive works that target solely finding the model that maximizes classification accuracy, the present work emphasizes methodological soundness, leakage-aware evaluation, and interpretability-driven auditing. To strive for this goal, this study proposes a Leakage-Aware Interpretability Framework (LAIF), which combines machine learning with SHAP-based model explainability, feature auditing, and domain knowledge to systematically identify potential leakage-prone variables and take measures.

We evaluated the proposed framework on two heterogeneous breast cancer datasets to study the impact of information leakage when predicting patient outcomes in clinical practice by simulating realistic clinical conditions with respect to predictive performance, feature importance, and model interpretability. The proposed framework adds a practical and reproducible dimension to incorporating explainability into the validation process to improve the reliability, transparency, and clinically relevant trust in an increasingly common layer of AI for oncology.

2. RELATED WORK

2.1 Machine learning in breast cancer prediction

The use of machine learning in breast cancer has been well established for diagnosis and prognosis. Commonly used models include logistic regression, SVM, RF, k-nearest neighbors (KNN), and gradient boosting methods like XGBoost. These methods had high discriminatory power between benign versus malignant cases and were also capable of being applied to in-patient stratification for recurrence prediction, using several evaluation metrics [2-4].

In addition to classification, machine learning has also been employed for more complex clinical prediction tasks. In the meantime, radiomics-based and deep learning models with incorporation of MRI and multimodal clinical features have been developed in order to predict lymphovascular invasion, treatment response, and recurrence risk [4, 17]. In contrast, comparative studies have shown a gradual increase in generalized predictions that are more powerful and less sensitive to instability when using ensemble- and tree-based learning methods within heterogeneous clinical settings [2, 17].

Recent works have instead tilted even more towards fusion of features from different data modalities (clinical, pathological, and imaging-derived information). Nevertheless, these multimodal approaches have significantly led to better patient stratification and show great promise to favor personalized treatment strategies in oncology [4, 5].

However, significant limitations remain despite these advances. Many studies exhibit limitations, including class imbalance, small sample sizes, absence of external validation,

and inconsistencies with preprocessing strategies that are not state-of-the-art [7, 8]. Many studies report predictive performance, but few assess fairness, which can lead to sources of potential methodological bias such as data leakage. This can affect reported performance and lead to an exaggerated idea of true predictive power.

Nowadays, additional attention to XAI methods is being given in most research studies due to the necessity of transparency and understanding of the internal reasoning process of machine learning models.

2.2 Explainable artificial intelligence in breast cancer research

The growing use of sophisticated machine learning models in healthcare has raised specific issues concerning transparency, interpretability, and clinical trustworthiness [13, 14]. XAI methods have been increasingly used in breast cancer applications to increase interpretability of model predictions and to facilitate clinically interpretable decision-making.

SHAP has become one of the most used methods for interpreting complex predictive models among existing XAI approaches [15]. SHAP delivers both global and local feature attributions by calculating the contribution of each feature to predictions made by a model. For medical applications, SHAP has been applied extensively for the discovery of important predictors, assessment of feature relevance, and determination of whether model behavior is consistent with meaningful clinical patterns [13, 18].

SHAP is available for breast cancer studies using different models based on clinical, pathological, radiomics, and multimodal datasets. Such analyses often identify tumor features, imaging-derived biomarkers, and patient-level clinical variables as prominent predictors, giving clinically meaningful information beyond traditional performance metrics [2, 4]. Subsequent works have integrated SHAP with deep learning and multimodal frameworks to study feature importance across heterogeneous patient populations [4, 17].

Despite these advances, XAI techniques still primarily function as post hoc interpreters of models after they have been created. Few studies have explored the potential for accessibility techniques to be used as a proactive auditing tool for identifying methodological failures (e.g., data leakage). Specifically, the possibility that highly influential SHAP features encode outcome-dependent information or temporal invalidity has been studied less [6, 10].

This highlights the need for more exploration into interactions of model interpretability, feature validity, and information leakage in clinical machine learning.

2.3 Data leakage and bias in medical machine learning

Data leakage is an important methodological problem for any machine learning model. It occurs when data not available at prediction time is inadvertently utilized in the construction or evaluation of a model, resulting in overly optimistic estimates of predictive accuracy [10].

Three main types of data leakage [10, 12] are target leakage, temporal leakage, and preprocessing leakage. Target Leakage refers to a condition where target information is captured by the variables directly and/or indirectly. Temporal leakage refers to the presence of data in the training set that was collected after the prediction time point. Preprocessing leakage occurs when processing steps (normalization, feature selection, imputation) are performed before the train-test splits. Out of these forms, target leakage is the most concerning within clinical datasets since we frequently have baseline patient information jointly with variables that are associated with the outcome [10, 18].

This leakage can be derived from several sources, like not selecting features properly, preprocessing before splitting the data, or introducing strong predictors to impute what we are going to predict. In the case of a clinical dataset, random variables such as survival status, follow-up outcomes, or post-treatment measurements directly encode the prediction label, which is why the predictive performance will be unrealistically high [6, 8].

The potential for leakage to allow near-optimal accuracy on models has been previously noted, with a corresponding significant reduction under proper experimental conditions [6]. Such observations highlight the need to be cognizant of both dataset design and audits of input variables and validation strategies in clinical machine learning research.

The leakage risk is also greater in breast cancer prediction studies since clinical datasets typically contain information on detailed follow-up and outcome-related variables. As an example, if survival indicators or relapse information are used as predictive inputs, they can inadvertently lead to bias. Despite this risk, the majority of studies do not state how leakage is avoided or assessed, raising concerns about methodological reliability and reproducibility [2, 6].

In addition, leakage may affect model interpretation as well. Explainability techniques may rank variables related to leakage as most impactful predictors, leading to erroneous clinical conclusions [13, 18]. However, the apparent interaction between leakage, interpretability, and theoretical applicability has not yet been sufficiently explored in the literature.

2.4 Research gap

Although earlier works have shown that machine learning and XAI methods can predict breast cancer, most of the current research treats interpretability as an independent methodological task from leakage prevention. Thus, little research has been conducted on the use of explainability methods as proactive evaluation techniques for the identification and resolution of data leakage in clinical machine learning systems.

This is especially relevant in breast cancer analysis datasets, which often contain outcome-related variables such as survival indicators, relapse data, and follow-up measurements presented side by side with the baseline clinical features. When such variables are introduced, it can result in over-inflated predictive performance, lower temporal validity, and more confusing clinical interpretations [6, 10].

Moreover, although SHAP has been widely used in the context of feature interpretation and model explainability, only a limited number of studies have systematically explored whether highly influential SHAP features could indicate leakage-prone information instead of clinically meaningful predictors [13, 18]. Consequently, the relationship between interpretability, validity of features, and robustness of methods is still largely under-researched in the literature.

Hence, leakage-aware frameworks that explicitly incorporate interpretability as well as feature auditing and validation procedures into the machine learning pipeline are required. To this end, the current work fills this gap by

conceptualizing a LAIF, which leverages SHAP not only for post hoc interpretation but also as part of a systematic mechanism to identify leakage-prone variables and evaluate predictive performance pre-and post-leakage mitigation.

3. METHODS

3.1 Datasets

In this study, we employed two publicly available breast cancer datasets that differ in nature to test the model performance under heterogeneous settings and to examine how data leakage may influence predictive models.

3.1.1 Mendeley dataset

The primary dataset for this study is from the Mendeley Data repository, which contains a large breast cancer cohort with demographic, pathological, molecular, treatment-related, and outcome-associated variables. Following data acquisition, the dataset consisted of 15,054 patients and 34 variables: a systematic aggregation of routine clinical information shared in most oncology studies. It contains patient characteristics, tumor descriptors, receptor status information, treatment indicators, as well as molecular subtypes and survival outcomes [19].

Relapse-free status was selected as the prediction target for binary prediction. In total, we included 8,976 patients labelled as Not Recurred and 6,072 patients labelled as Recurrence, a very imbalanced class dataset. This distribution is similar to that observed in clinical practice and requires an evaluation metric beyond conventional accuracy [20, 21].

Missing values: Preprocessing-stage median and mode imputation for numerical/categorical features was used to address missing observations in some variables. Since patients within a dataset remain unchanged and lose less information, these imputation methods are well-utilized in many clinical machine learning studies [22, 23].

One unique aspect of this dataset is that there are several follow-up and outcome variables, including overall survival status, overall survival (months), relapse-free status (months), and patient's vital status. However, predictive variables such as these are not present at the time of initial clinical decision-making, when intervention is to be guided. Therefore, using them as features when training a model can introduce target leakage and lead to overly optimistic performance estimation [2, 18].

The dataset is suitable for the purpose of demonstrating leakage detection and mitigation strategies within clinical machine learning as it contains several baseline clinical variables and information that is dependent on the outcome. This property of the dataset makes it especially appropriate for testing whether interpretability techniques can help not only detect potential leakage-inducing features but also increase the trustworthiness of predictive models.

A heterogeneous set of clinical, pathological, molecular, treatment-related, and outcome-associated variables from a large breast cancer cohort is summarized in Table 2, representing the Mendeley dataset. Furthermore, the existence of survival and follow-up features introduces a high-risk situation for target leakage, making it ideal in many respects to evaluate leakage detection and degradation methods within clinical machine learning.

A heterogeneous set of clinical, pathological, molecular, treatment-related, and outcome-associated variables from a large breast cancer cohort is summarized in Table 2, representing the Mendeley dataset. Furthermore, the existence of survival and follow-up features introduces a high-risk situation for target leakage, making it ideal in many respects to gauge both leak detection algorithms as well as degrading methods within clinical machine learning.

Table 2. Summary characteristics of the Mendeley breast cancer dataset used in this study

Characteristic	Value
Dataset source	Mendeley Data Repository
Number of patients	15,054
Number of variables	34
Target variable	Relapse-free status
Recurred cases	6,072
Non-recurrent cases	8,976
Data type	Clinical, pathological, molecular, treatment, and outcome-related data
Missing values	Present in several variables
Leakage risk	High
Examples of leakage-prone variables	Overall survival status, overall survival (months), relapse-free status (months), patient's vital status

3.1.2 Kaggle dataset

The Breast Cancer Wisconsin Diagnostic dataset from Kaggle was also used for benchmarking. The dataset contains 569 samples that are characterized using a total of 30 numerical features that have been computed from the digitized Fine Needle Aspiration (FNA) images [24].

This variable is known to be the target variable in tumor diagnosis, i.e., either malignant or benign. It is different from the Mendeley dataset, which provides not only baseline diagnostic measurements but also follow-up information, survival outcome, and post-treatment variables; whereas the Kaggle dataset contains only baseline information. Accordingly, the exercise with limited risk of target leakage from one dataset to another serves as a suitable reference dataset to validate our proposed modeling pipeline in conditions close to leakage-free [18].

Utilizing both datasets together allows for a synergistic evaluation approach, with the clean Kaggle benchmark allowing for initial testing of the model, and the Mendeley dataset providing real-world clinical challenges for external validation characterized by heterogeneous distributions across features, outcome-dependent variables, and high leakage risk [6, 10].

3.2 Data preprocessing

Both datasets underwent consistent preprocessing pipelines to ensure reproducibility and minimize methodological bias.

For numerical features, median imputation was used to fill in missing values, and the most common category for categorical variables [22, 23]. Only records with excessive missing values were filtered out from the analysis. Before transformation, features were divided into numerical and categorical types.

Z-score normalization was applied to numerical features. Crucially, scaling parameters were calculated on the training data only and then applied to the test set to prevent leakage of preprocessing [10, 12].

One-hot encoding was used to encode categorical variables in the clinical dataset. Finally, all preprocessing steps were

done inside a pipeline to make sure that transformations were only learned using the training data [12].

We used stratified sampling to split the datasets into training and testing subsets. To achieve an unbiased evaluation, the test set remained fully separated from all preprocessing and model training pipelines [21].

Prior to developing the models, class distribution was analyzed. The Mendeley dataset had mild class imbalance, where there were 8,976 non-recurred and 6,072 recurred cases. But without applying any synthetic oversampling techniques. However, model performance was evaluated using relevant and robust learning evaluation metrics, respectively: precision, recall, F1-score, and receiver operating characteristic-area under the curve (ROC-AUC), which are also more appropriate than accuracy for evaluating most moderately imbalanced classification models [20, 21].

3.3 Proposed Leakage-Aware Interpretability Framework

We introduce LAIF that seamlessly blends machine learning, interpretability, and feature auditing to systematically diagnose and prevent data leakage. The data preprocessing, predictive modeling, SHAP-based interpretation, domain-expert assessment, and leakage auditing are integrated into a unified validation workflow that is proposed to verify the model interpretability [2, 13, 18]. Figure 1 shows the general architecture of this framework.

Data preprocessing is the first step in the workflow, where different stages like handling missing values, encoding features, scaling features, and balancing classes are considered. We then organise the data into a training and test split to create an initial XGBoost classification model. The performance of models is evaluated using standard classification metrics like accuracy, precision, recall, F1-score, and ROC-AUC.

To enhance transparency and credibility, global and local interpretations of model behavior are presented using SHAP analysis. The development of a machine learning workflow usually involves model explainability being used as a post-exploitation process, but in the proposed framework, SHAP is integrated into the model validation. SHAP-based feature auditing is combined with domain-expert assessment of where information leakage [12, 13] might enter the data-generating process.

Features that were identified as being prone to leakage are forcibly audited and removed before the model is recreated. After removing leakage features, we re-trained the model using only clinically available baseline variables and then re-evaluated it to obtain leakage-free estimates of performance. This avoids spurious overfitting, thereby strengthening model robustness, while optimizing clinical generalizability of machine learning predictions [2, 4].

We present the LAIF framework, which enables a synergistic combination of explainability, domain knowledge, and leakage auditing to improve model transparency, bolster methodological rigor, and facilitate application development of clinically reliable machine learning systems.

3.4 Feature encoding

One-hot encoding was used to encode the categorical features of the clinical dataset to better capture their nominal nature [12].

The imaging dataset was not encoded since it contains only numerical features.

Over-and under-sampling methods were applied; this was because the pipeline works only on training data to avoid leakage, so it cannot be trained or fit on unseen data, but due to its design it does not suffer from any leakage [10, 12].

To facilitate SHAP interpretability analysis, feature names are kept across the entire pipeline, so that all encoded features can easily be mapped back to their clinical meaning [9, 18].

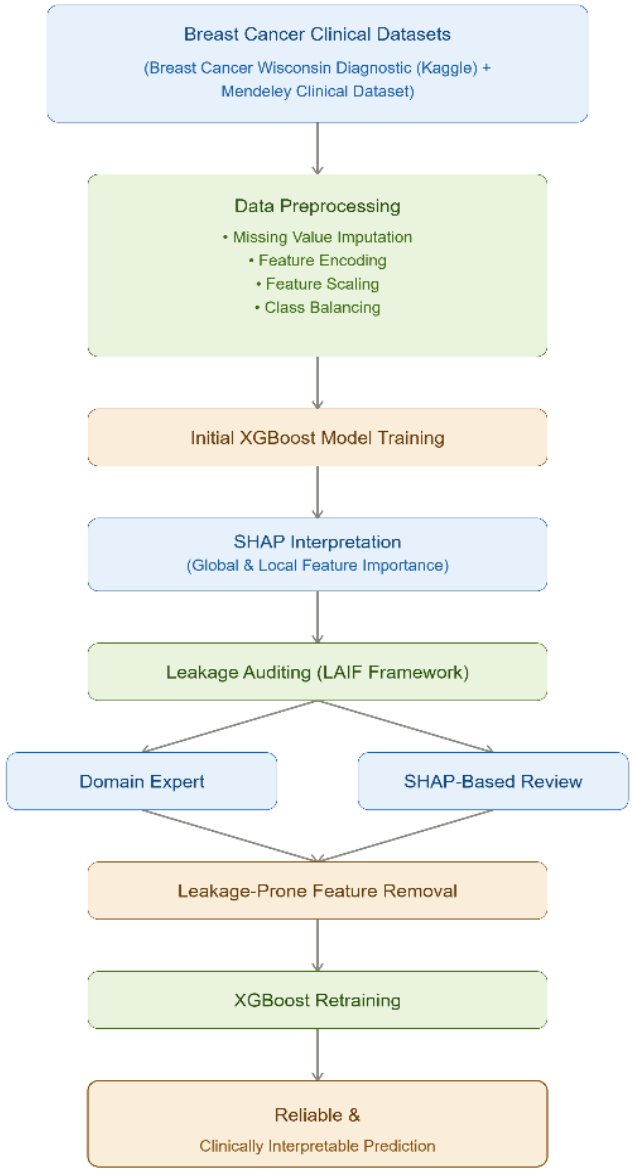


Figure 1. Proposed Leakage-Aware Interpretability Framework (LAIF) integrating SHapley Additive exPlanations (SHAP) analysis and domain knowledge for systematic leakage detection, feature auditing, model retraining, and leakage-free evaluation

3.5 Model development

Of the several classification algorithms tested, XGBoost was selected as our main classifier because it was selected from several classifiers due to its strong performance on structured clinical data with heterogeneous features and is used generally for structured clinical data with continuous and heterogeneous features [2, 4].

We trained the model using a gradient boosting framework, which sequentially fits decision trees to minimize prediction

error [3]. Hyperparameters, including the number of trees, maximum tree depth, and learning rate, were tuned using grid search with cross-validation on training data.

We applied regularization to avoid overfitting and gain better generalization, using techniques such as subsampling and depth constraints [12, 25]. Model selection occurred using only the ROC-AUC due to its robustness in evaluation compared to accuracy for moderately imbalanced classification problems [20, 21].

Other baseline models, including logistic regression and RF, were also tested for comparison.

3.6 Explainability analysis (SHapley Additive exPlanations)

Model predictions were interpreted using SHAP [15], and feature importance was analyzed. SHAP generates global and local explanations by approximating feature contributions to model output.

Global SHAP summary plots made it possible to see the most important features, and dependence plots allowed for analysis of how feature values affect predictions [15, 18].

SHAP was also used as a diagnostic tool to detect possible data leakage, in addition to its use as an interpretation tool. Any features with unusually strong influence were marked for follow-up investigation, notably outcomes-related features [6, 12].

3.7 Data leakage detection and mitigation

A systematic approach to identify and prevent data leakage was used. Initially, features were screened for clinical plausibility to identify features absent at prediction time. Second, SHAP analysis can be performed as a feature with excessively high importance [6, 12].

Early on, using all features gave very high predictive power, meaning there was probably unwanted target leakage. Outcome-related variables dominated model predictions, and this was further established with SHAP analysis [4].

We then removed these features and retrained the model with a clean dataset. We recreated the preprocessing pipeline, without using any information from the test set during training. This allowed pre- and post-cleaning model performance under leakage-aware conditions [12] to be compared without the need for environmental cleaning.

To mitigate potential leakage from unknown variables, a systematic review was done of all identified variables. Variables were categorized into those that would directly or indirectly cause leakage and eliminated if they leaked information unavailable at the planned prediction time [10, 26]. A summary of the leakage-prone variables and reasons for exclusion are presented in Table 3.

Some possible leakage features were detected by a two-stage auditing process. The first step was to study the SHAP summary plots, which enabled us to know which variables were especially emphasizing the prediction. The second step involved assessing the clinical focus of each top feature using domain knowledge to ascertain whether the associated information would be discernible at the time point intended for prediction. Those features that showed both high SHAP importance and contained information from either a post-outcome, follow-up, or survival-related point in time were labelled as leakage-prone variables and excluded from the final model [6, 10, 13]. This process made the leakage detection procedure more reproducible and transparent.

Performance dropped to more realistic but still robust levels after cleaning, suggesting that the model was identifying generalizable patterns in the data rather than leaking information [6, 26].

Table 3. Leakage-prone variables identified and removed from the Mendeley dataset [6, 10, 18]

Variable	Leakage Type	Reason for Removal
Overall survival status	Direct leakage	Contains post-follow-up outcome information strongly associated with recurrence status.
Overall survival (months)	Indirect leakage	Includes future survival information unavailable at prediction time.
Relapse-free status (months)	Direct leakage	Directly related to the target outcome and may reveal recurrence information.
Patient's vital status	Direct leakage	Reflects patient outcome after diagnosis and follow-up.

3.8 Evaluation metrics

Model performance was measured in terms of multiple classification metrics including accuracy, precision, recall, F1-score, and ROC-AUC.

While accuracy was reported, the focus was on recall and F1-score due to their clinical relevance in reducing false-negative predictions. This study used ROC-AUC as the main evaluation metric, which is a robust metric to evaluate moderately imbalanced classification [20, 21].

Moreover, precision–recall (PR) curves were also analyzed to offer an alternative assessment in the context of class imbalance [21].

We used stratified cross-validation for model training and evaluation to prevent information leakage between the training and testing data [10, 12]. The reproducibility of experimental setups was controlled by the use of fixed random seeds in stochastic methods [6, 12], and all experiments were implemented over fully reproducible pipelines.

4. RESULTS

4.1 Baseline performance on the Kaggle dataset

First, the proposed machine learning pipeline was tested on the Kaggle Breast Cancer Wisconsin dataset to build a leakage-free baseline. This dataset only includes baseline diagnostic measurements and is therefore not appropriate for the validation of the prediction framework under low-risk leakage conditions since no follow-up or outcome-related variables were available.

The Kaggle dataset showed a high classification accuracy and discriminative ability for the XGBoost classifier on all evaluation metrics.

The XGBoost model showed excellent predictive performance on the Kaggle dataset (Table 4). The model achieved perfect precision (1.0000) because there were no false positive predictions on the test set it was evaluated against. The very high ROC-AUC value (0.9947) showcases superior class separability and further verifies the robustness of the proposed leakage-free preprocessing and modeling

pipeline approach.

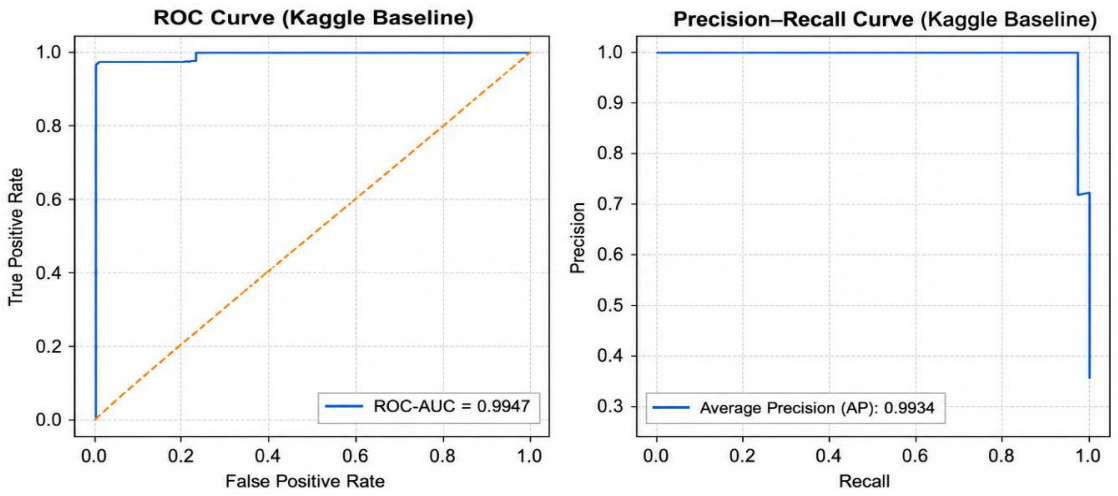
receiver operating characteristic (ROC) and PR curves were also derived to evaluate the model behavior further.

Figure 2(a) shows the ROC curve close to the upper-left corner and excellent class separability between benign and malignant samples. Also, based on Figure 2(b), where all the recalls are below ~0.8, and we still achieve high precision again confirms that the invariant behavior leakage-free baseline model is robust and consistent.

Table 4. Performance evaluation of the Extreme Gradient Boosting (XGBoost) model on the Kaggle Breast Cancer Wisconsin dataset

Dataset	*n*_samples	*n*_features	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Kaggle	569	30	0.9649	1.0000	0.9048	0.9500	0.9947

Note: ROC-AUC = receiver operating characteristic-area under the curve.



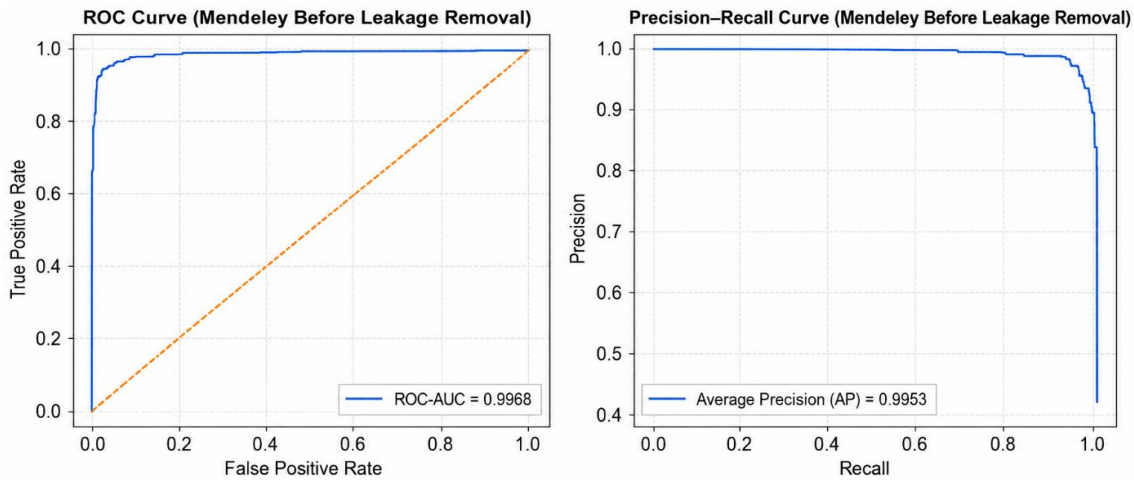
(a) (b)

Figure 2. Receiver operating characteristic (ROC) and precision–recall (PR) performance curves for the Kaggle Breast Cancer Wisconsin dataset: (a) ROC curve demonstrating excellent discriminative capability of the Extreme Gradient Boosting (XGBoost) classifier (receiver operating characteristic-area under the curve (ROC-AUC) = 0.9947); (b) PR curve showing stable precision across recall thresholds

Table 5. Performance evaluation of the Extreme Gradient Boosting (XGBoost) model on the Mendeley dataset before leakage removal

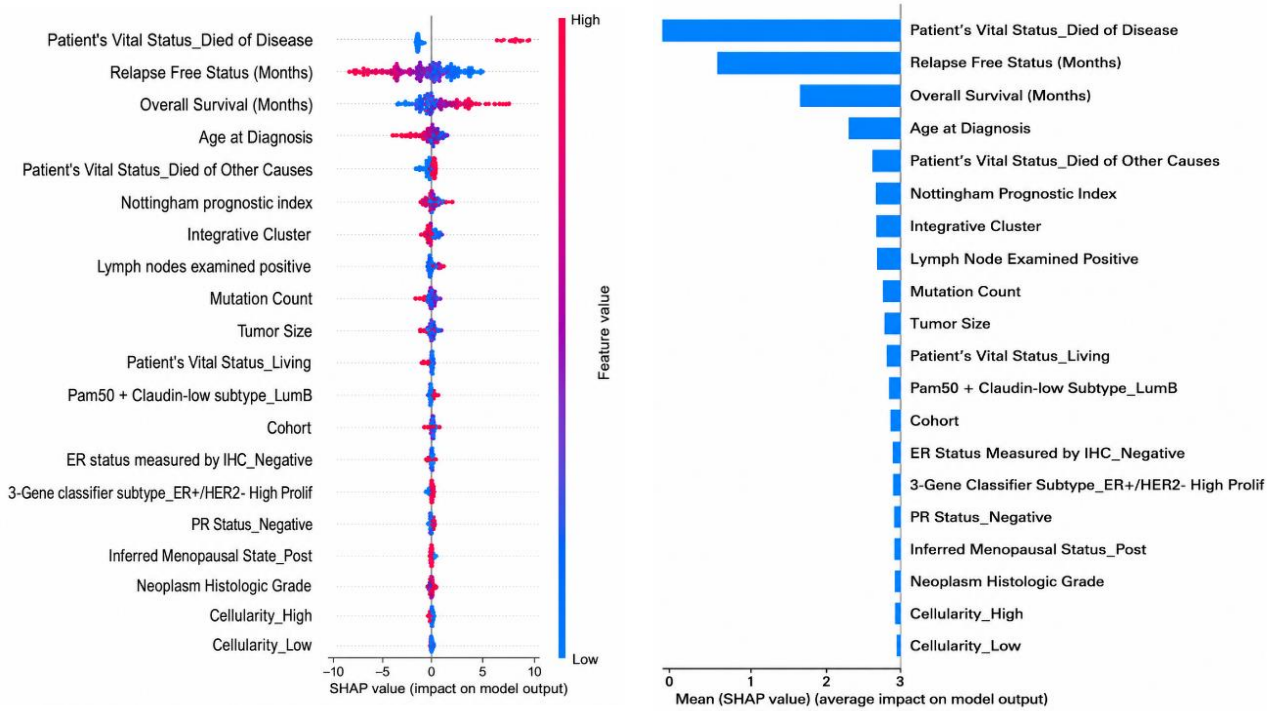
Dataset	*n*_samples	*n*_features	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Mendeley before leakage removal	5,000	91	0.9710	0.9770	0.9504	0.9635	0.9968

Note: ROC-AUC = receiver operating characteristic-area under the curve.



(a) (b)

Figure 3. Receiver operating characteristic (ROC) and precision–recall (PR) performance curves for the Mendeley dataset before leakage removal: (a) ROC curve demonstrating exceptionally strong discriminative capability of the Extreme Gradient Boosting (XGBoost) classifier (receiver operating characteristic-area under the curve (ROC-AUC) = 0.9968); (b) PR curve showing consistently high precision across recall thresholds



(a) (b)

Figure 4. SHapley Additive exPlanations (SHAP)-based interpretability analysis before leakage removal on the Mendeley dataset: (a) SHAP summary plot demonstrating the influence of outcome-related variables on model predictions; (b) SHAP feature importance bar plot showing the dominance of leakage-prone variables prior to mitigation

4.2 Performance on the Mendeley dataset before leakage removal

After the baseline was validated on top of the Kaggle dataset, the proposed framework was applied to the Mendeley clinical breast cancer dataset with leakage auditing turned off and using the full set of features. The Mendeley dataset, on the other hand, is more heterogeneous (clinical, pathological, molecular, treatment-related features, and outcome variables), better corresponding to clinical prediction scenarios compared with the Kaggle dataset.

The results before any leakage removal are very strong indeed, with scores from every metric being in the high ranges of possible values for predictions using XGBoost.

As can be seen in Table 5, the predictive performance of this model was extremely strong before leakage mitigation, with an accuracy score of 0.9710 and a ROC-AUC of nearly perfect, which is 0.9968. Meanwhile, precision, recall, and F1-score also stayed approximately constant as well, signaling a good discriminative ability. Since the dataset includes survival-related and follow-up variables, we performed additional interpretation-based auditing to verify that the observed performance was not influenced by information leakage. Mendeley dataset ROC and PR curves for the Mendeley dataset to visualize classification behavior prior to leakage mitigation.

As shown in Figure 3(a), the ROC curve nears the vertical axis in the upper-left corner, which clearly shows very strong class separability before leakage mitigation. Likewise, as shown in Figure 3(b), the precision remained high over a range of recall values. While this might seem exceptionally strong predictive performance at first glance, the ROC-AUC and PR behavior were well beyond what was thought to be achievable without data leakage on a dataset.

4.3 SHapley Additive exPlanations-based leakage detection

Before commencing with leakage mitigation, the Mendeley dataset was analyzed using SHAP-based interpretability analysis to understand model behavior and reveal variables that may have been impacted by leakage.

Depending on the output variable, Figure 4(a) demonstrates an exaggerated value of SHAP importance for some outcome-dependent variables. There were also some very important variables contributing to model predictions; these include patient's vital status, overall survival (months), and relapse-free status (months). As shown in Figure 4(b), these variables made a significant contribution to the classification prediction of several patient samples.

These variables are all derived from future outcome information, which means it is a clear target leakage situation that would not be possible to obtain at the intended time of prediction in real-world clinical settings. The dominant contribution of these variables implies that some portion of the predictive performance was achieved by reading future outcome information rather than true prognostic learning from baseline clinical characteristics.

After this analysis, all potential leakage variables were systematically eliminated before retraining the model in a leakage-aware manner.

4.4 Performance after leakage removal

Leakage detection implemented excluded from the dataset all variables with any type of post-outcome, follow-up, or survival information. Subsequently, we built and ran a new pipeline from preprocessing to model training on clinically available baseline variables only.

The resulting cleaned model showed more clinically

plausible predictive performance, albeit low compared to pre-leakage.

Table 6. Comparative performance of the Extreme Gradient Boosting (XGBoost) model before and after leakage removal

Model Version	Accuracy	Precision	Recall	F1-score	ROC-AUC
Before leakage removal	0.9710	0.9770	0.9504	0.9635	0.9968
After leakage removal	0.8950	0.9049	0.8263	0.8638	0.9416

Note: ROC-AUC = receiver operating characteristic-area under the curve.

However, after removing variables with the potential to leak information into training labels (see Table 6), performance deteriorated. Accuracy is now down from 0.9710 to 0.8950; ROC-AUC has decreased from .9968 to.9416. A similar decrease was found for precision, recall, and F1-score.

This drop demonstrates that most likely the predictive performance was to some degree inflated because of data leakage. Nonetheless, the leakage-free model showed good discriminative performance, suggesting baseline variables meaningful in the clinic still harbor prognostic signal.

4.5 Post-leakage SHapley Additive exPlanations interpretation

SHAP analysis was repeated after leakage removal to assess the reliance of the retrained model on clinically plausible predictors.

Following leakage mitigation, outcome-related variables no longer appeared among the highest-ranked predictors Figure 5. Rather, simple clinically relevant variables including Age at Diagnosis, Tumor Size, Lymph Nodes Examined Positive, Mutation Count, and Nottingham Prognostic Index were the most significant predictors in their respective models.

In contrast, the baseline variables in the SHAP value distribution show a more uniform contribution to multiple baseline variables, suggesting that the retrained model learned clinically relevant prognostic patterns and did not depend on future outcome information. Moreover, the increased distribution of SHAP values over patients indicates better generalization and more realistic clinical decision-making under leakage-leave conditions.

4.6 Comparative performance summary

For an overall comparison of model performance before and after leakage mitigation, we plotted the respective performance across all evaluation metrics.

As shown in Figure 6, predictive performance on all evaluation metrics dropped steadily after removing leakage. The accuracy dropped from 0.9710 to 0.8950, and ROC-AUC decreased from 0.9968 to 0.9416. The same reduction could be observed for precision, recall, and F1-score.

The observations therefore corroborate the concern that a portion of the original predictive power is an artefact of information in leakage-prone variables that is dependent on the outcome. Importantly, the leakage-free model maintained strong discriminatory ability, suggesting that clinically meaningful baseline covariates continue to provide predictive information on breast cancer prognosis.

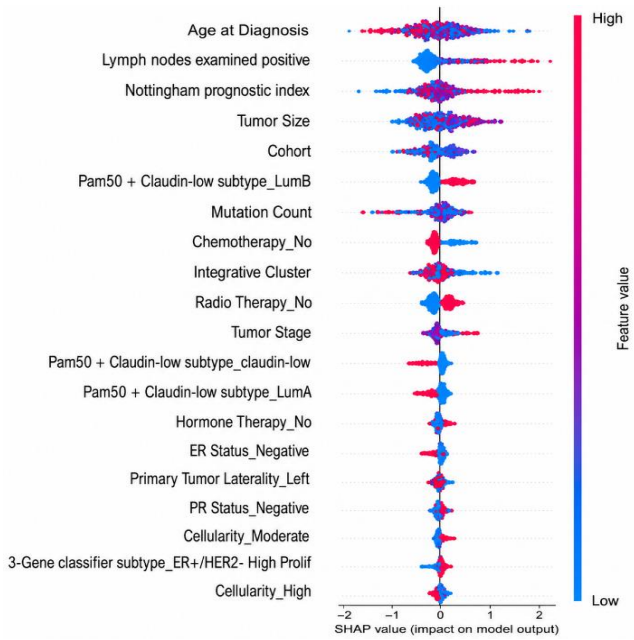


Figure 5. SHapley Additive exPlanations (SHAP) summary (beeswarm) plot after leakage removal, demonstrating that the retrained model relies primarily on clinically relevant baseline variables rather than outcome-dependent information

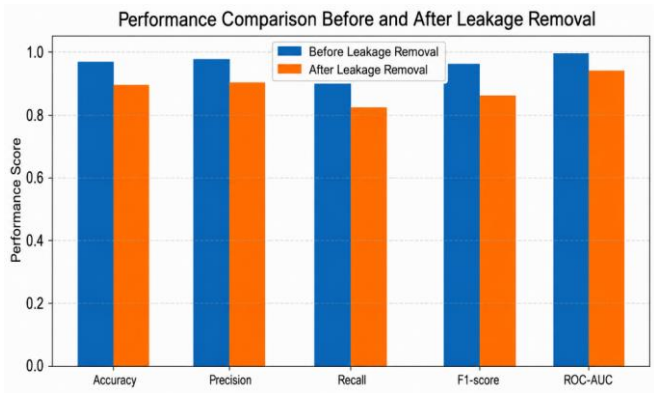


Figure 6. Comparative performance of the Extreme Gradient Boosting (XGBoost) model before and after leakage removal across accuracy

4.7 Stability analysis and computational cost

The leakage-free XGBoost model was trained and evaluated across a range of random seeds to assess the robustness of modeling premises. Furthermore, we measured computational efficiency through training and inference times.

In Table 7, the leakage-free XGBoost model stability is shown over repeated experimental runs. Despite moderate variability in recall values, the relatively low standard deviation of both ROC-AUC and F1-score suggests stable classification across different random train–test splits.

The decline in recall when compared to the primary experiment reflects the added modeling challenge posed by constraining feature selection to clinically relevant baseline variables and changing sample initialization conditions. However, the model exhibited reasonable predictive stability and robustness across multiple evaluations.

The presented framework not only allows for efficient training and inference times but also suggests that the

approach is computationally feasible to use in repetitive experimental evaluations and feasible to include in clinical machine learning workflows.

Table 7. Stability and computational efficiency analysis of the leakage-free Extreme Gradient Boosting (XGBoost) model across multiple random seeds (mean ± standard deviation)

Metric	Mean ± Standard Deviation
Accuracy	0.8224 ± 0.0045
Precision	0.8610 ± 0.0237
Recall	0.6680 ± 0.0180
F1-score	0.7519 ± 0.0057
ROC-AUC	0.8980 ± 0.0020
Training time (s)	0.0943 ± 0.0152
Inference time (s)	0.0070 ± 0.0001

Note: ROC-AUC = receiver operating characteristic-area under the curve.

4.8 Additional model comparison

More importantly, to explore why leakage-aware evaluation raised different responses to various machine learning algorithms, we evaluated some classification models post-leakage removal. We contrast linear, kernel-based, boosting, and ensemble learning methods to determine if leakage has an impact on different families of models.

Post-leakage removal, XGBoost and RF exhibited the best performance across all metrics, as seen in Table 8. XGBoost showed the most favorable trade-off between sensitivity and discrimination among all evaluated models, followed closely by RF with slightly elevated ROC-AUC.

While CatBoost obtained competitive precision, its recall performance was lower than other classifiers, which suggests it has the least sensitivity to identify recurrent cases under leakage-free conditions. On the other hand, SVM was relatively weak across most of the evaluation metrics, indicating SVM's limited capacity to capture various complex nonlinear relationships from heterogeneous clinical data.

These results show that leakage-aware evaluation has an impact on a variety of machine learning algorithms and reinforce the robustness of the proposed auditing framework across model families. Furthermore, the high performance of tree-based ensemble methods indicates that it is beneficial to consider nonlinear feature effects for breast cancer outcome prediction.

Table 8. Performance comparison of different machine learning models after leakage removal

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
RF	0.9180	0.9373	0.8536	0.8935	0.9831
SVM	0.7330	0.6675	0.6725	0.6700	0.8092
CatBoost	0.7980	0.8407	0.6154	0.7106	0.8918
XGBoost	0.9240	0.9291	0.8784	0.9031	0.9521

Note: ROC-AUC = receiver operating characteristic-area under the curve; RF = random forest; SVM = support vector machines; XGBoost = Extreme Gradient Boosting.

5. DISCUSSION

We show that data leakage encodes highly misleading signals which may distort clinical machine-learning models, stressing the importance of interpretability-driven validation to ultimately yield reliable and clinically relevant results. The results indicate that, based on the assessment of model behavior across two heterogeneous datasets, our proposed XGBoost framework is relatively robust to distributional shifts in training data, outcome-dependent feature contamination poses considerable risk.

The testing results on the Kaggle dataset show a strong leakage-free baseline. High predictive power is obtained by using strictly defined imaging-derived features, and this shows that the proposed preprocessing and modelling pipeline can capture informative signal in leakage-free conditions. This trend is also corroborated by the corresponding SHAP analysis, establishing that the model utilized clinically justified imaging morphology features such as tumor size and texture-related features. When the importance conferred by the data and our understanding of its process are aligned, we gain confidence in both the validity and interpretability of our model.

Compared to the initial experiments on the Mendeley dataset, which clearly overfitted and reached unprecedented predictions, with predictions very close to 1s (not realistic for a real-world clinical prediction task). That level of performance showed that leakage is occurring, rather than indicating a more capable model. Incorporation of domain knowledge and SHAP-based explanations revealed outcome-dependent variables as explanatory factors; survival status and relapse-driving characteristics proved to be dominant drivers of model predictions. This would be a clear scenario of target leakage, and the worst methodological pitfall imaginable, even if you do not have these variables when making predictions (which, of course, represents one of the canons on which machine learning methodology relies).

The model performance is therefore more reasonable after eliminating features that are related to leakage. The importance of the reduction is not due to model degradation; rather, it shows that shortcut learning has been removed, and clinically valid prediction behavior has been adopted. Importantly, the model retained reasonable discrimination, suggesting that baseline clinical characteristics carry some meaningful prognostic information. This result lends additional evidence to the resource-inefficient premise that, for many contexts with suitable validation processes, good prediction is achievable with outcome-derived variables.

From a methodological perspective, this work shows how SHAP interpretation methods can aid in both explaining model predictions and conducting audits of dataset integrity. Detection of leakage even in complex clinical datasets cannot be fully addressed by traditional evaluation metrics. This work presents a practical and reproducible methodology, primarily because its validation can be incorporated into existing explainability pipelines to uncover and mitigate hidden sources of bias.

On a clinical front, the implications highlight how deploying models trained on tainted data can be perilous. One consequence of such leakage is that models with poor generalization ability might appear to perform well in development but would subsequently underperform in real-world settings, leading to inappropriate conclusions about risk and clinical decisions. To mitigate this, we propose a leakage-aware framework to ensure that predictions are only made based on features that can be observed at the intended point of use, making them more reliable and leading to safer clinical decision-making.

This work integrates SHAP as a central model validation

step within the modeling pipeline, in contrast to many previous studies that considered explainability a post hoc visualization step. This method allows for systematic leakage identification and better understanding of model behavior. While breast cancer datasets are used in this current analysis, the framework proposed here is applicable to other clinical disciplines that involve longitudinal or outcome-rich data.

In summary, our work provides a practical and reproducible approach to robustness-aware machine learning with interpretability as part of the model validation process instead of providing an optional post hoc analysis.

6. LIMITATIONS

Despite the rigorous methodology of this study, several limitations should be acknowledged.

Two publicly available datasets with limited heterogeneity but obtained from actual clinical environments were utilized for the first analysis. Future work should include enhanced generalizability through external validation with independent, multi-institutional datasets.

Another point is that the leak detection was achieved using a combination of domain knowledge and SHAP-based interpretability analysis. Although this method has worked, it is still somewhat reliant on expert judgment. The automation of leakage detection frameworks could improve scalability and objectivity, especially in high-dimensional datasets.

Third, this study used one machine learning algorithm (XGBoost). While this approach may be ideal for structured clinical data, further work can evaluate whether similar leakage effects and mitigation results are seen in other modeling approaches (such as deep learning or ensemble methods).

Fourth, there was no explicit implementation of temporal validation by splitting strategies for longitudinal data. In addition, using time-aware validation schemes would help to further bolster the robustness of model evaluation, especially for future clinical use.

Lastly, interpretability analysis was done using SHAP values mainly. Although SHAP is a powerful and well-established method based on theory, the addition of further interpretability methods could result in an even more extensive view of the model's behavior.

Despite these limitations, the study presents a pragmatic and reproducible approach to leakage-aware machine learning and extensive insights into ways to increase the fidelity and clinical applicability of prediction models.

7. CONCLUSION

This study also illustrates how data leakage can invalidate a clinical machine learning model and the importance of validation that is driven by interpretability for reliable performance assessment. Using an XGBoost-based framework, we demonstrate that near-perfect predictive performance of survival can arise from the unintended inclusion of outcome-dependent features rather than true prognostic learning when evaluated on two heterogeneous breast cancer datasets.

Using systematic guided leakage detection coupled with SHAP-based analyses, we identified and removed outcome-related variables from the feature list and achieved improved leak-free performance characteristics that are measured in ways that better approximate clinical use. After removing leakage, predictive metrics were lowered, and the model kept high discriminative power, indicating that baseline clinical features encode real predictive signal.

This study demonstrates that it shows that explainability methods can not only be used as model interpretability tools, but also as effective ways to identify methodological weaknesses such as data leakage.

The proposed framework provides guidance on how to develop clinically relevant prediction models by ensuring that performance reflects data available at the time of prediction.

This work emphasizes the necessity of leakage-aware evaluation and interpretability-centered validation for clinical machine learning, leading to more reliable and clinically deployable AI systems.

AUTHOR CONTRIBUTIONS

Conceptualization, M.H. and K.A.M.H.; methodology, M.H. and K.A.M.H.; validation, A.F.A. and A.F.M.; formal analysis, M.H.; investigation, M.H.; resources, K.A.M.H., A.F.A., and A.F.M.; data curation, M.H. and K.A.M.H.; writing—original draft preparation, M.H.; writing—review and editing, K.A.M.H., A.F.A., and A.F.M.; visualization, M.H.; supervision, A.F.A. and A.F.M.; project administration, A.F.A. All authors have read and agreed to the published version of the manuscript.

DATA AVAILABILITY

The data used to support the findings of this study are available from the corresponding author upon request.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

REFERENCES

[1] Sung, H., Ferlay, J., Siegel, R.L., et al. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians, 71(3): 209-249. https://doi.org/10.3322/caac.21660

[2] Ma, M.W., Liu, R.Y., Wen, C.J., et al. (2022). Predicting the molecular subtype of breast cancer and identifying interpretable imaging features using machine learning algorithms. European Radiology, 32(3): 1652-1662. https://doi.org/10.1007/s00330-021-08271-4

[3] Sidey-Gibbons, J.A.M., Sidey-Gibbons, C.J. (2019). Machine learning in medicine: A practical introduction. BMC Medical Research Methodology, 19: 64. https://doi.org/10.1186/s12874-019-0681-4

[4] Wu, X.Y., Liu, H., Liu, J.Y., et al. (2026). Multimodal deep learning with routine clinical data for recurrence risk stratification in HR+/HER2− early breast cancer. Research, 9: 1136. https://doi.org/10.34133/research.1136

[5] Esteva, A., Robicquet, A., Ramsundar, B., et al. (2019).

A guide to deep learning in healthcare. Nature Medicine, 25: 24-29. https://doi.org/10.1038/s41591-018-0316-z

[6] Kapoor, S., Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. Patterns, 4(9): 100804. https://doi.org/10.1016/j.patter.2023.100804

[7] Kelly, C.J., Karthikesalingam, A., Suleyman, M., Corrado, G., King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine, 17: 195. https://doi.org/10.1186/s12916-019-1426-2

[8] Roberts, M., Driggs, D., Thorpe, M., et al. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nature Machine Intelligence, 3: 199-217. https://doi.org/10.1038/s42256-021-00307-0

[9] Lekadir, K., Frangi, A.F., Porras, A.R., et al. (2025). FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ, 388: e081554. https://doi.org/10.1136/bmj-2024-081554

[10] Kaufman, S., Rosset, S., Perlich, C., Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. ACM Transactions on Knowledge Discovery from Data, 6(4): 15. https://doi.org/10.1145/2382577.2382579

[11] Goldstein, B.A., Navar, A.M., Pencina, M.J., Ioannidis, J.P.A. (2017). Opportunities and challenges in developing risk prediction models with electronic health records data: A systematic review. Journal of the American Medical Informatics Association, 24(1): 198-208. https://doi.org/10.1093/jamia/ocw042

[12] Kuhn, M., Johnson, K. (2013). Applied Predictive Modeling. Springer. https://doi.org/10.1007/978-1-4614-6849-3

[13] ElShawi, R., Sherif, Y., Al-Mallah, M., Sakr, S. (2021). Interpretability in healthcare: A comparative study of local machine learning interpretability techniques. Computational Intelligence, 37(4): 1633-1650. https://doi.org/10.1111/coin.12410

[14] Samek, W., Wiegand, T., Müller, K.R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. ITU Journal: ICT Discoveries, 39-48. https://www.itu.int/dms_pub/itu-s/opb/journal/S-JOURNAL-ICTF.VOL1-2018-1-P05-PDF-E.pdf.

[15] Lundberg, S.M., Lee, S.I. (2017). A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA, pp. 4768-4777. https://dl.acm.org/doi/10.5555/3295222.3295230

[16] Collins, G.S., Moons, K.G.M., Dhiman, P., et al. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ, 385: e078378. https://doi.org/10.1136/bmj-2023-078378

[17] Li, J.P., Qiu, Z.X., Cao, K.Y., et al. (2023). Predicting muscle invasion in bladder cancer based on MRI: A comparison of radiomics, and single-task and multi-task deep learning. Computer Methods and Programs in Biomedicine, 233: 107466. https://doi.org/10.1016/j.cmpb.2023.107466

[18] Molnar, C. (2019). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Leanpub. https://originalstatic.aminer.cn/misc/pdf/Molnar-interpretable-machine-learning_compressed.pdf.

[19] Pereira, B., Chin, S.F., Rueda, O.M., et al. (2016). The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. Nature Communications, 7: 11479. https://doi.org/10.1038/ncomms11479

[20] Chicco, D., Jurman, G. (2020). Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Medical Informatics and Decision Making, 20: 16. https://doi.org/10.1186/s12911-020-1023-5

[21] He, H.B., Garcia, E.A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9): 1263-1284. https://doi.org/10.1109/TKDE.2008.239

[22] Little, R., Rubin, D. (2019). Statistical Analysis with Missing Data. Hoboken, NJ, USA: John Wiley & Sons. https://doi.org/10.1002/9781119482260

[23] Stekhoven, D.J., Bühlmann, P. (2012). MissForest—Non-parametric missing value imputation for mixed-type data. Bioinformatics, 28(1): 112-118. https://doi.org/10.1093/bioinformatics/btr597

[24] Street, W.N., Wolberg, W.H., Mangasarian, O.L. (1993). Nuclear feature extraction for breast tumor diagnosis. Biomedical Image Processing and Biomedical Visualization, 1905: 861-870. https://doi.org/10.1117/12.148698

[25] Ying, X. (2019). An overview of overfitting and its solutions. Journal of Physics: Conference Series, 1168(2): 022022. https://doi.org/10.1088/1742-6596/1168/2/022022

[26] Roberts, D.R., Bahn, V., Ciuti, S., et al. (2016). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography, 40(8): 913-929. https://doi.org/10.1111/ecog.02881