











Comparison of Explainable Artificial Intelligence Model and Radiologist Review Performances to Detect Breast Cancer in 752 Patients

Pelin Seher Oztekin, MD , Oguzhan Katar, MSc , Tulay Omma, MD , Serap Erel, MD ,
Oguzhan Tokur, MD , Derya Avci, PhD , Murat Aydogan, PhD , Ozal Yildirim, PhD ,
Engin Avci, PhD , U. Rajendra Acharya, PhD, DEng, DSc 

Received May 17, 2024, from the Department of Radiology, University of Health Sciences, Ankara Training and Research Hospital, Ankara, Turkey (P.S.O., O.T.); Department of Software Engineering, Firat University, Elazig, Turkey (O.K., M.A., O.Y., E.A.); Department of Endocrinology and Metabolism, University of Health Sciences, Ankara Training and Research Hospital, Ankara, Turkey (T.O.); Department of Surgery, University of Health Sciences, Ankara Training and Research Hospital, Ankara, Turkey (S.E.); Department of Computer Technology, Firat University, Elazig, Turkey (D.A.); School of Mathematics, Physics, and Computing, University of Southern Queensland, Springfield, Queensland, Australia (U.R.A.); and Centre for Health Research, University of Southern Queensland, Springfield, Queensland, Australia (U.R.A.). Manuscript accepted for publication July 13, 2024.

We would like to thank to Firat University Scientific Research Projects Coordination Unit (FÜBAP) for their support. This work has been supported by FÜBAP under project number ADEP.23.21.

Address correspondence to Ozal Yildirim, Department of Software Engineering, Firat University, Elazig, Turkey.

E-mail: ozalyildirim@firat.edu.tr

Abbreviations

ACR, American College of Radiology; BI-RADS, Breast Imaging Reporting and Data System; DT, decision trees; K-NN, k-nearest neighbor; LIME, Local Interpretable Model-agnostic Explanations; LR, logistic regression; MRI, magnetic resonance imaging; PRF, pulse repetition frequency; RF, random forest; SHAP, SHapley Additive exPlanations; SR, strain ratio; SVM, support vector machine; US, Ultrasonography; X2GAI, explainable XGBoost model; XAI, Explainable Artificial Intelligence; XGBoost, Extreme Gradient Boosting

doi:10.1002/jum.16535

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Objectives—Breast cancer is a type of cancer caused by the uncontrolled growth of cells in the breast tissue. In a few cases, erroneous diagnosis of breast cancer by specialists and unnecessary biopsies can lead to various negative consequences. In some cases, radiologic examinations or clinical findings may raise the suspicion of breast cancer, but subsequent detailed evaluations may not confirm cancer. In addition to causing unnecessary anxiety and stress to patients, such diagnosis can also lead to unnecessary biopsy procedures, which are painful, expensive, and prone to misdiagnosis. Therefore, there is a need for the development of more accurate and reliable methods for breast cancer diagnosis.

Methods—In this study, we proposed an artificial intelligence (AI)-based method for automatically classifying breast solid mass lesions as benign vs malignant. In this study, a new breast cancer dataset (Breast-XD) was created with 791 solid mass lesions belonging to 752 different patients aged 18 to 85 years, which were examined by experienced radiologists between 2017 and 2022.

Results—Six classifiers, support vector machine (SVM), K-nearest neighbor (K-NN), random forest (RF), decision tree (DT), logistic regression (LR), and XGBoost, were trained on the training samples of the Breast-XD dataset. Then, each classifier made predictions on 159 test data that it had not seen before. The highest classification result was obtained using the explainable XGBoost model (X²GAI) with an accuracy of 94.34%. An explainable structure is also implemented to build the reliability of the developed model.

Conclusions—The results obtained by radiologists and the X²GAI model were compared according to the diagnosis obtained from the biopsy. It was observed that our developed model performed well in cases where experienced radiologists gave false positive results.

Key Words—breast cancer; explainable AI; machine learning; ultrasound

Breast cancer remains a major health problem worldwide and is the fourth leading cause of cancer-related deaths worldwide, after lung, liver, and stomach cancer.¹ The gold standard method for breast cancer screening is mammography.² Ultrasonography (US) is an indispensable complement to mammography, especially in women with dense and extremely dense breast structures. It is also the first diagnostic method used to evaluate women under 40 years of age with average risk.³ The

US is preferred because it is radiation-free, non-invasive, widely used, affordable, and can be easily used in invasive procedures. The addition of Doppler and elastographic evaluations to grayscale imaging in routine practice has increased the diagnostic accuracy of US examinations.⁴ The Breast Imaging Reporting and Data System (BI-RADS) Atlas developed by the American College of Radiology (ACR) provides a standardized approach to the evaluation and management of breast lesions.⁵ The BI-RADS-US evaluation is subjective in the absence of a clear clinical decision rule and is highly dependent on the reader's experience.⁶ The low specificity and high false positive results due to both the imaging method and the reporting system used to result in unnecessary biopsies and follow-ups, causing unnecessary anxiety and wasted time and resources. Therefore, a new problem-solving method should be investigated to improve diagnostic performance in the evaluation of breast lesions. Research has turned to developing machine learning-based automatic classifier models to overcome the limitations of manual analysis processes.⁷ These models take radiological images as input, extract various features, and use these features to distinguish lesions as benign or malignant.

Machine learning is an approach that allows computer systems to learn through data-driven experiences. Image-based detection methods are widely used in computerized breast cancer diagnosis.^{8,9} Hand-crafted features extracted from images obtained from imaging techniques such as mammography, US, magnetic resonance imaging (MRI), and so on, are used for classical machine learning techniques.^{10–13} The main difficulty of these approaches is that the feature extraction phase requires expertise and experience in image processing. In machine learning, deep learning approaches have recently provided an end-to-end learning structure by eliminating the difficulties of hand-crafted feature extraction.¹⁴ Thus, studies on breast cancer detection on images using deep learning approaches have gained momentum.^{15–17} Although image data is an important parameter in breast cancer detection, it is not sufficient by itself. For this reason, creating clinic datasets by expert radiologists using various tests and features obtained from the image is a useful resource. The Wisconsin dataset¹⁸ was published in 1993 and has been widely used in ML-based breast cancer detection.^{19–21} With

the development of medical testing and imaging methods, clinical datasets with up-to-date features can enable more reliable studies in this field.

While ML-based classifiers are highly accurate in healthcare studies, there is an important need for the classification process to be explainable. The inherent black-box nature of some classifiers leads to difficulties in understanding the specific features they prioritize for classifications.²² This lack of transparency hinders understanding model decisions, leading to concerns about reliability and suitability.²³ Explainable artificial intelligence (XAI) approaches have been developed in response to these issues. XAI approaches are designed to increase the ability to understand and explain decision mechanisms within complex machine-learning models.²⁴ In this context, algorithms such as SHAP (SHapley Additive exPlanations)²⁵ and LIME (Local Interpretable Model-agnostic Explanations)²⁶ are popular methods, especially for understanding why classifiers make certain predictions.

In this study, we proposed an explainable classifier system for automatically detecting benign and malignant breast solid mass lesions on one-dimensional (1D) data. We used a novel, huge private patient dataset to train and validate the classifier. Preprocessing steps were applied to the dataset samples and six different classifiers were trained on the same training samples. Using the weights obtained from the training, the performance of each classifier was validated on the same test data. The results of the most successful model were statistically compared with the diagnosis of an expert with 23 years of experience. The features that the model focuses on in its predictions are presented with SHAP and LIME methods and compared with the features the expert focuses on in his diagnosis. Furthermore, to the best of our knowledge, this is the first study in which the focus features of a classifier trained on a 1D breast cancer dataset are described and compared with the focus features of an experienced expert's predictions.

The main contributions of this study can be summarized as follows:

- Presented a new breast cancer dataset (Breast-XD), including data from expert radiologists from comprehensive and up-to-date medical technologies.

This dataset is shared publicly for research purposes.

- Employed classifiers and achieved high success rates in breast cancer detection.
- For clinical applications, comprehensive analyses of the results obtained by artificial intelligence models are performed and interpreted by radiologists.
- To overcome the disadvantages of black-box approaches regarding reliability, an explainable model (X²GAI) was created with SHAP and LIME approaches.
- The differences and similarities between the results obtained by the X²GAI model and the diagnoses of expert radiologists are analyzed and discussed in detail.

Materials and Methods

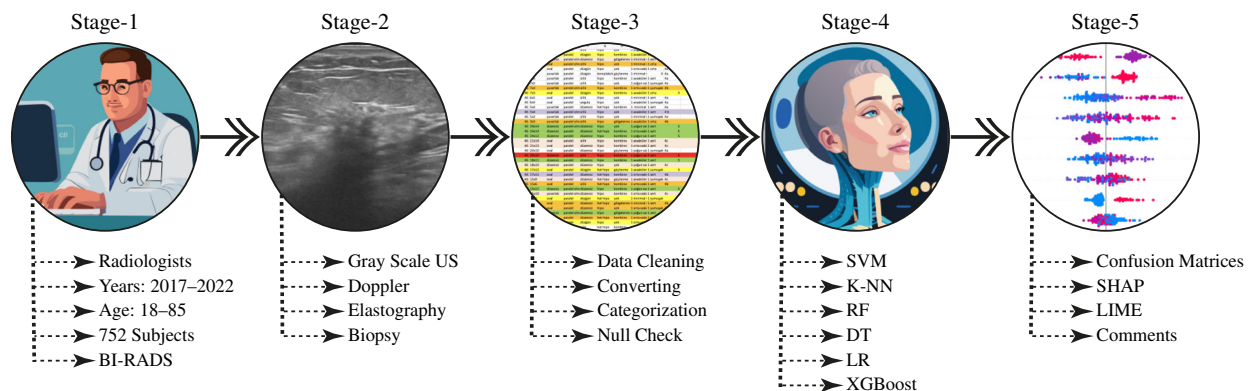
This study presents an explainable AI-based approach to breast cancer detection. The clinical data used in the study was obtained from subjects who were admitted to the hospital or followed up with suspected breast cancer over 5 years. A unique dataset was created using nine different visual features determined by radiologists specialized in breast cancer diagnosis. Various pre-processing steps were applied to this data and transformed into a format suitable for evaluating AI models. Hence, six different machine

learning models are trained using the dataset samples, and their performance is evaluated on a separate test set. The most successful model is selected for comparison with an experienced expert. To gain insight into the decision-making process of the model, the features it focused on during its predictions are visualized using SHAP and LIME methods. The general structure of the proposed method is shown in Figure 1.

Breast-XD Dataset

In this study, a special dataset was created with the approval of the Ankara Training and Research Hospital Clinical Research Ethics Committee (E-93471371-514.99, E-22-1150). This retrospective study was planned to include 1000 patients aged 18 to 85 years who underwent breast US examination (grayscale, Doppler, and elastography) and core biopsy with BI-RADS 3, 4 (a, b, c), and 5 between January 2017 and November 2022. However, 248 patients were excluded from the study for various reasons. Some did not have their gray-scale US, Doppler US, and/or US-elastography images registered in the system. Others did not receive a definitive diagnosis due to the absence of core biopsy or excision procedures. Additionally, those with benign diagnoses who did not undergo follow-up were excluded. The Breast-XD dataset was created with 791 solid mass lesions from 752 different patients (dataset available at: <https://kaggle.com/datasets/zalyildirim/breast-cancer-dataset>). All lesions included in the dataset had core biopsy

Figure 1. The steps of this study: *Stage-1 Data Collection:* Data on individuals with suspected breast cancer or who were followed up over 5 years were compiled by experts. *Stage-2 Creating Dataset:* Creating the dataset by extracting nine different visual features on the values obtained from radiological tests for each subject. *Stage-3 Pre-Processing:* Data is converted into formats suitable for the inputs of AI models. *Stage-4 Model Training:* Training of AI models and performance evaluation with various metrics. *Stage-5 Evaluation:* Determination of the features that the classifier AI models effectively concentrate on in the decision phase (SHAP and LIME) and radiologist evaluations.



procedures performed using 14-gauge fully automatic 14-gauge sharp needles and sufficient specimens were obtained (3–6 specimens). The steps each patient went through while creating the Breast-XD dataset and the features obtained from these steps are given in Figure 2.

The first column of the Breast-XD dataset contains the age of the included patients. The youngest patient is 18 years old and the oldest is 85 years old, with an average age of 49. The distribution of patient ages in the Breast-XD dataset based on biopsy results is given in Figure 3. The most important factor that draws attention to the figure is that the proportion of patients with malignant tumors increases as the age parameter increases.

The Breast-XD dataset was created by a breast radiologist with 23 years of experience in breast radiology using a Hitachi High-Vision Preirus (Hitachi Medical Corp, Tokyo, Japan) with a linear transducer (50 mm, 13–6 MHz). The ultrasound images collected for the study were selected from the sections with optimal morphologic assessment of the lesion. The length, width, shape, orientation, margin, echo pattern and posterior features of the lesions obtained from these slices were added to the feature columns of the dataset, respectively. The distribution of lesion

shape and margin features based on biopsy results is shown in Figure 4.

The number of subjects with oval-shaped lesions was higher than the others and the biopsy results of oval-shaped lesions were found to be benign at a high rate. On the other hand, lesions with irregular shapes were mostly diagnosed as malignant. Similarly, most cases with irregular margin distribution were classified as malignant.

This study analyzed Doppler US images to characterize the presence of blood vessels and blood flow within the lesions. Doppler images were created from the slices with the highest amount of signal with appropriate pulse repetition frequency (PRF) and gain settings to prevent artifact formation. Strain ratio (SR), a semi-quantitative method, was used to extract other features that could not be obtained from grayscale and Doppler images of the lesions. Ten features obtained from three different radiological imaging methods represent the input samples of the Breast-XD dataset, while the biopsy results of the patients represent the class label, which is the feature of the dataset to be predicted. The features of the Breast-XD dataset, their short descriptions, and the values or value ranges they can take are presented in Table 1.

Figure 2. An illustration of the creation of the Breast-XD dataset.

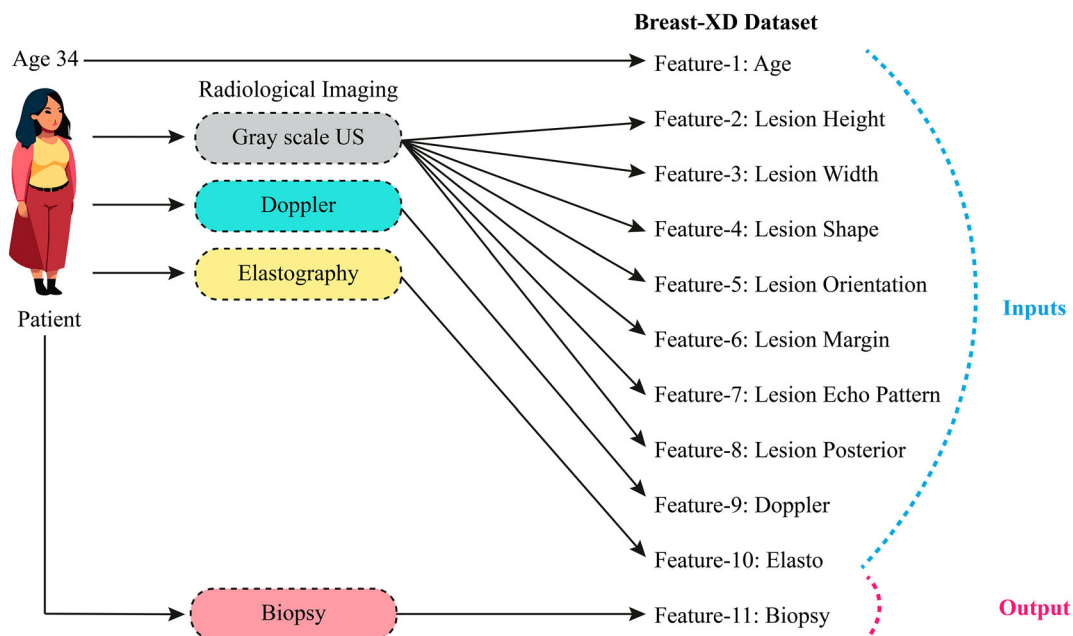


Figure 3. Distribution of subjects in Breast-XD dataset according to biopsy results.

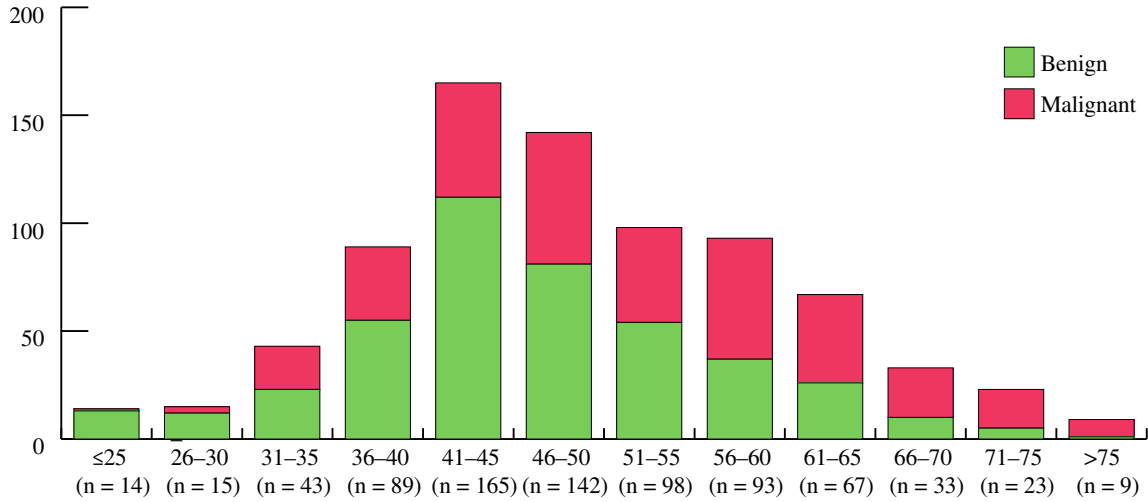
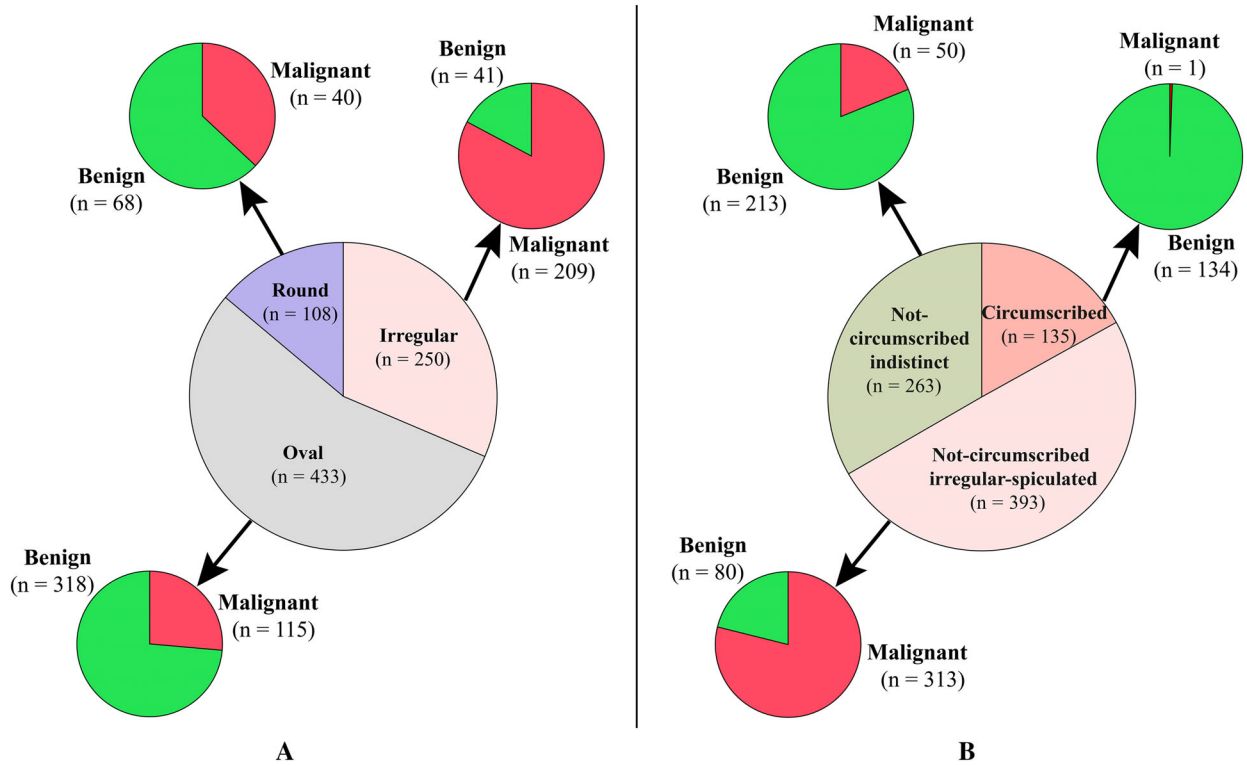


Figure 4. Class distribution of cases in the dataset according to lesion shape and margin characteristics: (A) Shape distribution; (B) Margin distribution.



Machine Learning-Based Classifiers

Machine learning classifiers classify data into different categories or classes and are usually trained by

supervised learning methods.²⁷ Classifiers can predict or classify future data by feeding the training with data. This study uses six machine learning classifiers

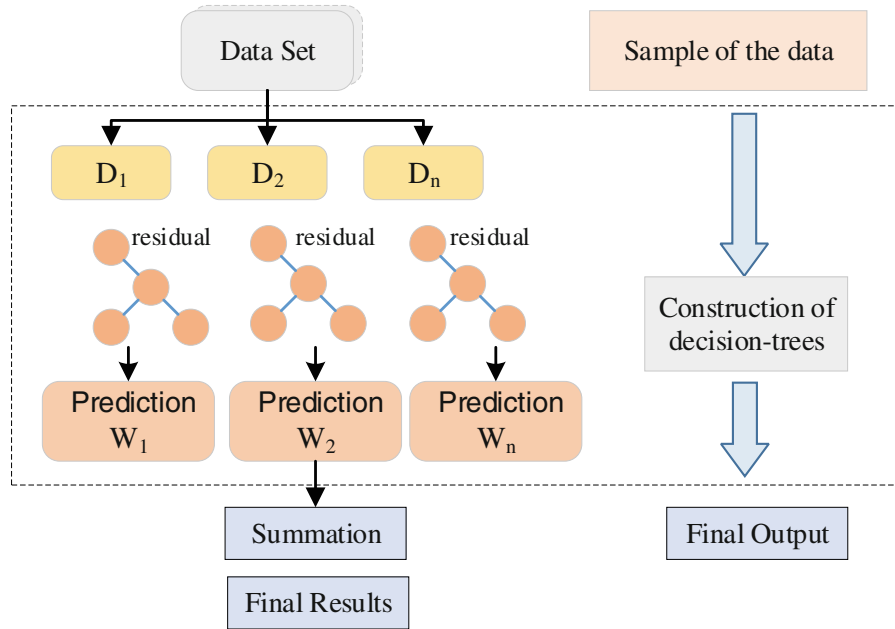
Table 1. Attributes and Value Ranges of Records in the Breast-XD Dataset

Features	Description	Type	Values
Patient age	Represents the age of the patient.	Numeric	18–85
Lesion height	The length of the lesion was measured in mm.	Numeric	3–75
Lesion width	The width of the lesion was measured in mm.	Numeric	3–55
Lesion shape	Represents the shape of the lesion.	Label	Irregular (0), Oval (1), and Round (2)
Lesion orientation	Represents the orientation of the lesion.	Label	Parallel (0) and non-parallel (1)
Lesion margin	Refers to the margin of the lesion.	Label	Circumcibred (0), not-circumscribed irregular-spiculated (1), not-circumscribed indistinct (2)
Lesion echo pattern	Describes the echogenicity of the lesion.	Label	Iso (0), Complex (1), and Hypo (2)
Lesion posterior	Refers to the posterior features of the lesion.	Label	No (0), Acoustic (1), Combined (2), and Shadow (3)
Doppler	Refers to the results of the Color Doppler US examination of the lesion.	Label	Avascular (0), Minimal Vascular (1), Moderate Vascular (2), and Dense Vascular (3)
Elasto	Explain the results of Compressive Elastography.	Label	Soft (0), Medium (1), and Hard (2)
Biopsy	Shows biopsy results.	Label	Benign (0) and Malignant (1)

to classify benign and malignant lesions for breast cancer detection. The workings of these classifiers are briefly summarized as follows.

1. *Support vector machine (SVM)* is a simple, powerful, and efficient supervised algorithm for solving classification and regression problems.²⁸ It creates a discriminative optimal hyperplane that takes a low-dimensional input vector and maps it to a higher-dimensional feature space to provide a high generalization network capability.²⁹ The concept of an optimal separating hyperplane can be used in both cases, where the data is linearly separable and non-linearly separable.
2. *Decision trees (DT)* are algorithms that build the tree structure from the top to classify the data in the first stage. The tree structure is named root, branches, and leaves, starting from the top. Branches are connected to nodes, with each branch connected to the root at the top. Each attribute in the data represents a node in the tree after classification.³⁰ This tree structure works by partitioning the dataset based on certain features and using a set of decision rules to make the best decision in each partition.
3. *The k-nearest neighbor (K-NN)* algorithm is a machine learning algorithm used in a classification or regression problem. This algorithm calculates the k-nearest neighbors (k) around data points, using a numerical distance metric.³¹ For classification, the classes of the k-nearest neighbors around a sample are examined and the most frequent class is assigned as the prediction.³² For regression, a prediction value is calculated using the values of the k-nearest neighbor's target variables
4. *Logistic regression (LR)* is a statistical classification method. It predicts which class a data point belongs to between two or more classes. This method is used to model the relationship of a dependent variable (output class) with independent variables. Logistic regression gives the results as probability values and then classifies them based on a given cut-off point. Logistic regression is based on probability calculations using an S-curve called a logistic function, which helps to solve classification problems.³³
5. *Random forest (RF)* is an efficient machine-learning algorithm that consists of traditional decision tree classifiers. A bootstrap bagging technique is applied to generate training subsets for each tree. The classifier's output is usually decided by a majority voting technique for each tree and is considered the cumulative decision of each tree.³⁴
6. *XGBoost* is a powerful classification and regression algorithm called "Extreme Gradient Boosting." This algorithm is based on tree-based learning methods and often successfully performs high-performance classification tasks on large and complex datasets. XGBoost builds a strong predictor by combining many weak predictors, reducing the

Figure 5. A block representation shows the XGBoost classifier’s working structure.³⁵



overfitting problem.³⁵ It can also be optimized to improve the model’s performance using regularization terms and customizable loss functions. The structure of the XGBoost classifier is given in Figure 5.³⁵

XGBoost is a method that has many advantages over common methods due to its high predictive power, its ability to prevent overlearning, its ability to manage null data, and its speed.³⁶ XGBoost, optimized especially for working with large datasets, is seen as more advantageous than other methods due to its ability to regularize, prune, work with null values, and optimize the system. The basic design of XGBoost has the following objective function as in Equation 1.³⁷

$$L^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \quad (1)$$

In Equation 1, l is the loss function to minimize errors, t is the number of iterations, $\Omega(f_t)$ is the additional regularization term for model complexity, y_i is the observed value, \hat{y}_i is the predicted value calculated by the equation. Assuming the model has k decision trees, the predicted value is in Equation 2.³⁷

$$\hat{y}_i = \sum_{f \in F} f_k(x_i), f_k \in F \quad (2)$$

In Equation 2, F is the set of regression trees and f is a regression tree in the set. The regularization element defines the complexity of the tree and Equation 3 increases the stability of the model by continuously simplifying it.

$$\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \quad (3)$$

In Equation 3, γ and λ are the regularization parameters, w is the score of each leaf, and T is the number of leaves of the tree.

Explainable Artificial Intelligence

AI classifiers work like black boxes. When learning data, classifiers automatically adjust many parameters and it is unclear how these parameters work. Therefore, explaining why classifiers make certain decisions or how they make predictions is difficult. This black box structure can complicate measuring the model’s performance and assessing its reliability.²² However, explainability methods such as SHAP and LIME have recently been used to overcome the “black box”

nature of machine learning classifiers that complicate understandability and reliability.^{38,39}

SHAP is a game theory-based explainability method that uses Shapley values to calculate how much each attribute contributes to a prediction.⁴⁰ Shapley value is a concept that helps determine how players should share rewards based on their contributions in a cooperative game. Adapting this concept to a machine learning context, players become data features, while the reward represents a predicted value. The mathematical equation used to calculate the Shapley value is given in Equation 4.⁴¹

$$\phi_i(f) = \frac{1}{N!} \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \tag{4}$$

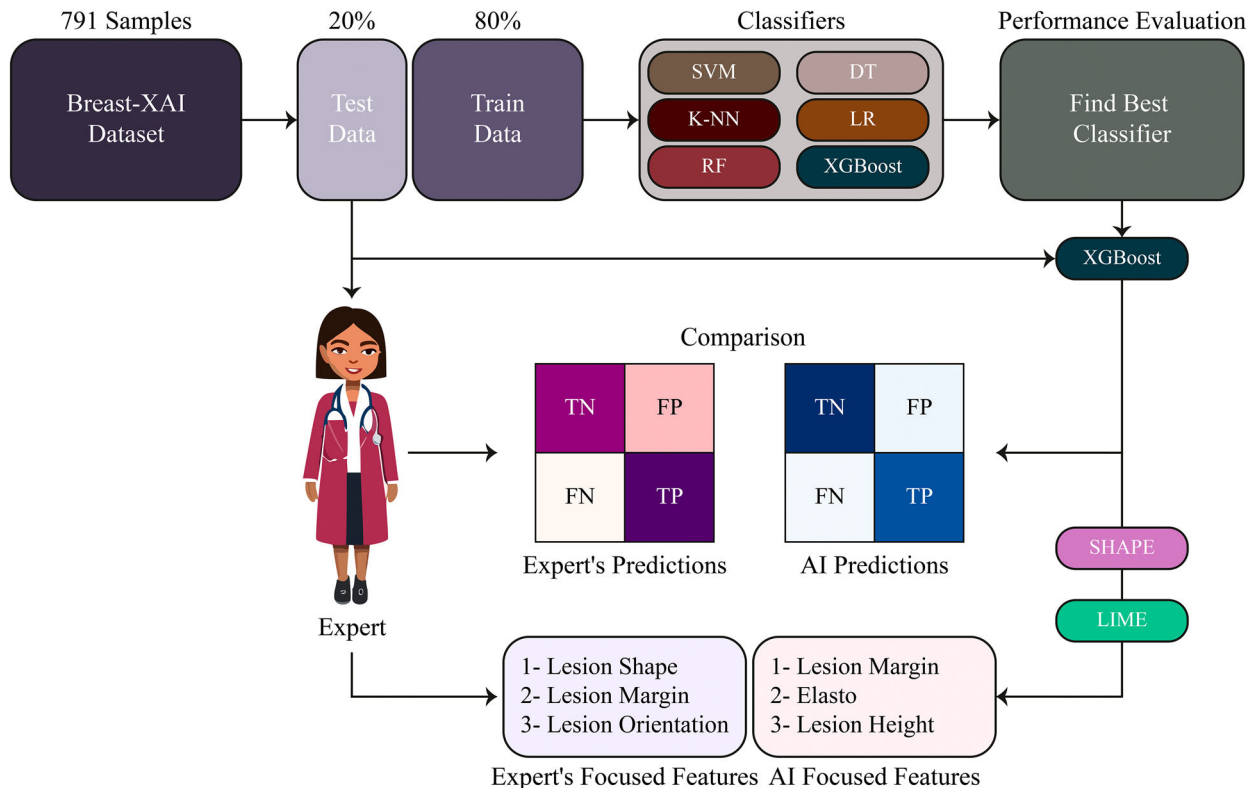
In this equation, f is the prediction function, N is the total number of players, S is the number of player groups, $f(S)$ is the prediction result of player group S , $f(S \cup \{i\})$ is the prediction result of adding a player i

to player group S . Each player i is added to all player groups, and the change in the prediction result as a result of this addition is calculated. The average value of these changes represents the Shapley value of the player.

LIME produces a meaningful, simplified explanation using sample data to explain how the model classifies a given sample.⁴² This method is widely used to understand the internal logic of black-box models and to make the model's decisions interpretable by humans. The working steps of the LIME algorithm can be summarized as follows.

1. *Data Perturbation*: LIME takes the sample data that needs to be explained and creates new examples by slightly modifying this data. These new samples will be used to better understand the model's behavior.
2. *Prediction*: The perturbed instances are evaluated on the model and the model's predictions about each perturbed instance are obtained. These predictions represent a prediction distribution for the original sample.

Figure 6. Block representation of the experimental method used in this study.



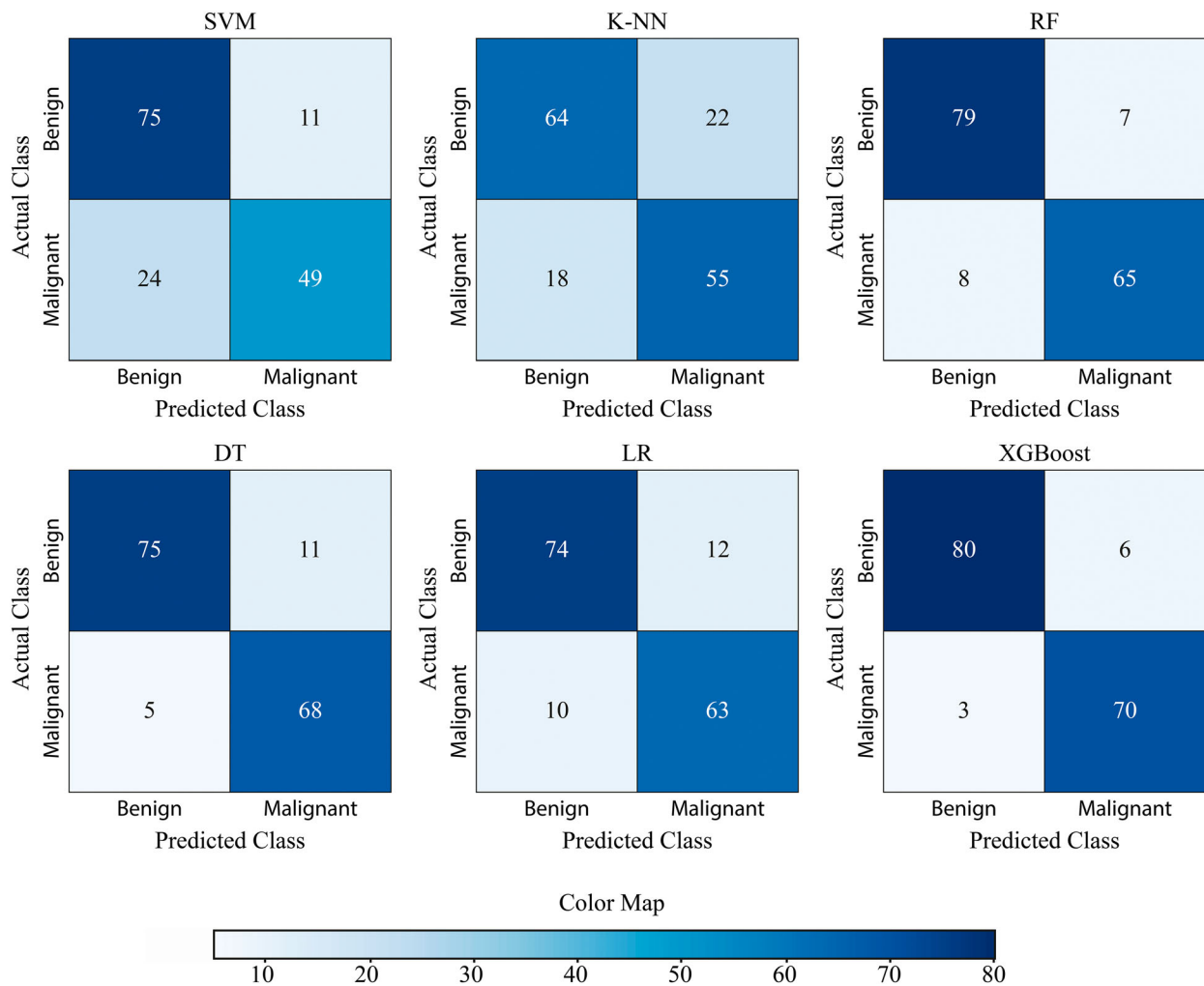
3. *Feature Selection*: It applies a weighting process to identify the features that effectively classify the instance that needs to be annotated. This is done to understand how sensitive the model is to which features.
4. *Model Fitting*: The weighted samples build a local linear model based on the identified influential features. This local model represents the local behavior of the complex model.
5. *Explanation*: The generated local model can be explained in a form that is better understood by humans. This explanation shows the important features that influence the model’s decisions and how these features interact.

These two methods make the predictions of complex classifiers more understandable and interpretable, making it easier to measure the model’s performance and assess its reliability. At the same time, these explainability techniques can also be used to detect erroneous or misleading predictions of the model, which can improve the reliability of the model.

Experimental

In this study, six classifiers were trained and tested on the Breast-XD dataset. In the training phase, 80% of the dataset samples and 20% of the dataset samples

Figure 7. Confusion matrices containing test predictions of machine learning models.



in the testing phase were used. The most successful classifier was determined by comparing the performance of the classifiers on the test samples. Then, the features focused on the most successful classifier were

visualized with SHAP and LIME algorithms. A comparison of the machine learning classifier with the predictions of expert radiologists and the focused features in breast cancer detection is presented. A block

Figure 8. ROC curves obtained using various classifiers.

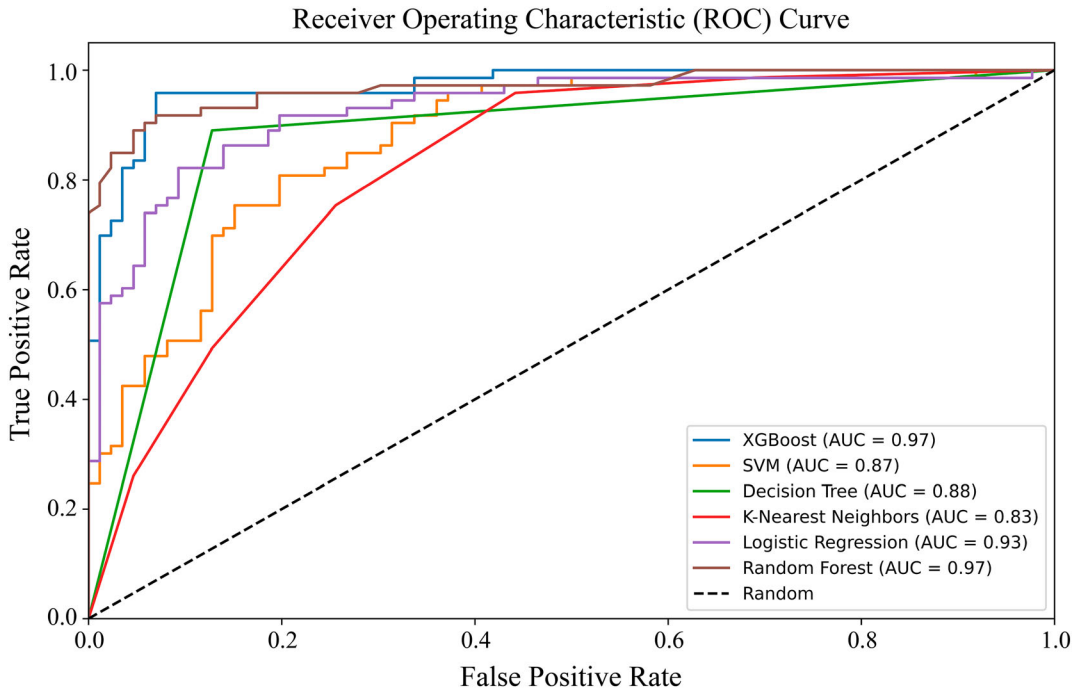


Figure 9. Confusion matrix (left) and ROC curve (right) obtained based on the test samples estimated by the radiologist.

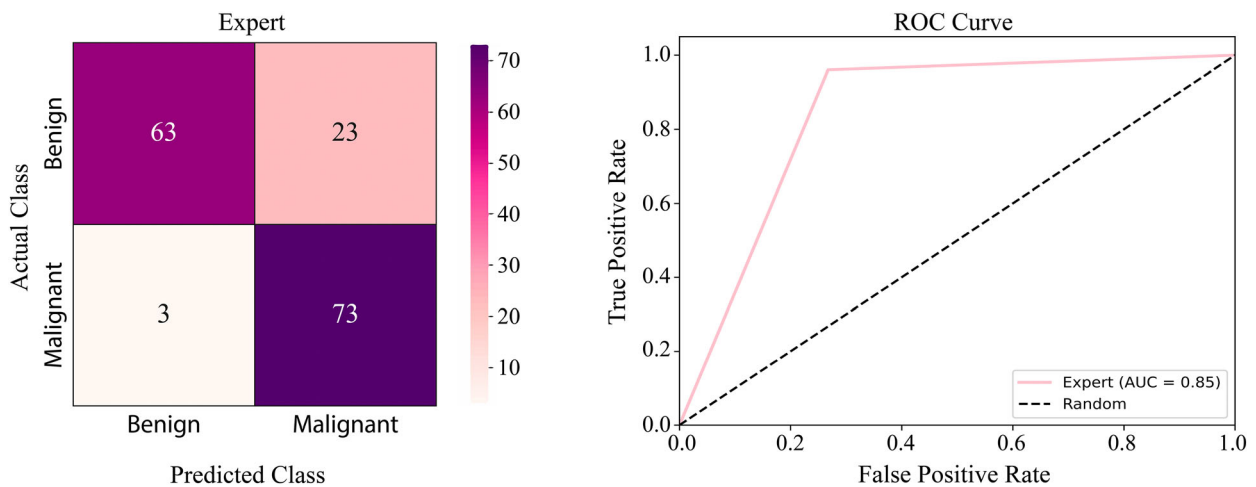


diagram representation of the experimental method used in this study is given in Figure 6.

Experimental Setups

This study used six classifiers to classify breast lesions as benign or malignant using the Breast-XD dataset. In the training phase, all classifiers used the default hyper-parameters of the scikit-learn library. The training of the classifiers is performed on 632 samples randomly selected from the Breast-XD dataset samples and kept constant for all classifiers. Then, the classifiers were tested on 159 samples not seen during the

training phase. The performance of the classifiers is compared using confusion matrix-based statistical metrics. The predictions of the most successful model are then characterized using SHAP and LIME methods. All experimental studies were conducted in the Google Colab environment with the Python programming language version 3.10.12 on the CPU processing unit.

Experimental Results

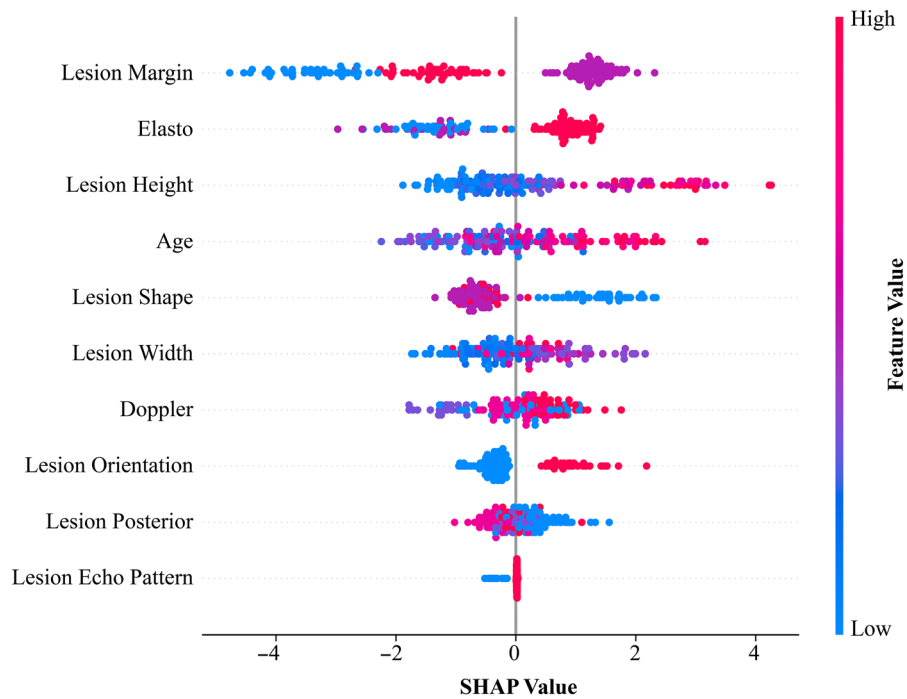
Six classifiers, SVM, K-NN, RF, DT, LR, and XGBoost, were trained on the training samples of the

Table 2. Performances Obtained Using the AI Models and Radiologists on the Test Data

Classifier	Number of False Predictions	Accuracy (%)	Precision (%)	Recall (%)	F-1 Score (%)
SVM	35	77.99	81.67	67.12	73.68
K-NN	40	74.84	71.43	75.34	73.33
RF	15	90.57	90.28	89.04	89.65
DT	16	89.94	86.07	93.15	89.47
LR	22	86.16	84.00	86.30	85.14
XGBoost	9	94.34	92.11	95.89	93.96
Radiologist	26	83.95	76.04	96.05	84.88

Note: The highest values are marked in bold.

Figure 10. Features and SHAP values that the X²GA model focused on in the test set.



Breast-XD dataset. Then, each classifier made predictions on 159 test data that it had not seen before. The confusion matrices obtained are given in Figure 7.

The test samples' most and least successful classifiers are XGBoost and K-NN, respectively. Our

results show that XGBoost is the best-performing classifier, with only 6 misclassifications among 86 benign labels and three misclassified instances among 73 malignant labels. The ROC curves obtained by all classifiers in the test phase are given in Figure 8. It may be noted that all classifiers achieved

Figure 11. Prioritization of the features considered by radiologists and X²GAI (the numbers next to each feature represent the order of priority).

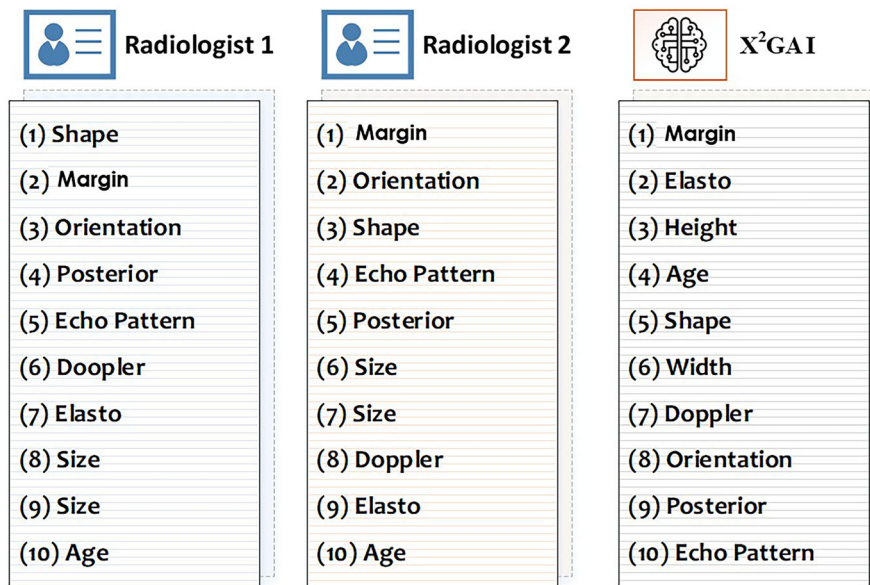


Figure 12. A 53-year-old patient had a hypochoic mass with an irregular shape and indistinct margin, measuring 10 × 7 mm in size at 10 o'clock in the right breast with a combined posterior feature and no parallel orientation. Doppler US showed no significant vascular flow through the lesion. US-Elastography shows that the lesion has a hard structure. The lesion was evaluated as highly suspicious for malignancy with its radiologic features.



an AUC rate of above 80%. However, the classifier with the highest AUC is XGBoost.

Radiologist Versus AI Models

To benchmark the machine learning model’s ability to classify benign and malignant breast lesions, a radiologist with 23 years of experience was asked to classify the same test samples. The confusion matrix and ROC curve generated based on the radiologist’s diagnosis of the test samples are presented in Figure 9.

The experienced radiologist performed 23 false positive classifications on 86 benign labeled test data.

Among the 76 samples labeled malignant, there were only three false negatives. The experimental studies obtained using six different machine learning classifiers and the experienced radiologist on the Breast-XD dataset samples are presented in Table 2.

Explainable Model X²GAI

It may be noted from Table 2 that most of the classifiers have performed better than the experienced radiologist. However, the lack of explanation of the features that AI focuses on in its predictions causes such studies not to be accepted by experts in clinical

Figure 13. In the US image of a 32-year-old female subject, there is a 14 × 11 mm, non-parallel oriented, hypoechoic, irregular not-circumscribed lesion with posterior shadowing at 12 o’clock in the left breast. Doppler US showed minimal vascular flow through the lesion. US-Elastography shows that the lesion has a hard structure. The lesion was evaluated as highly suspicious for malignancy with its radiologic features.

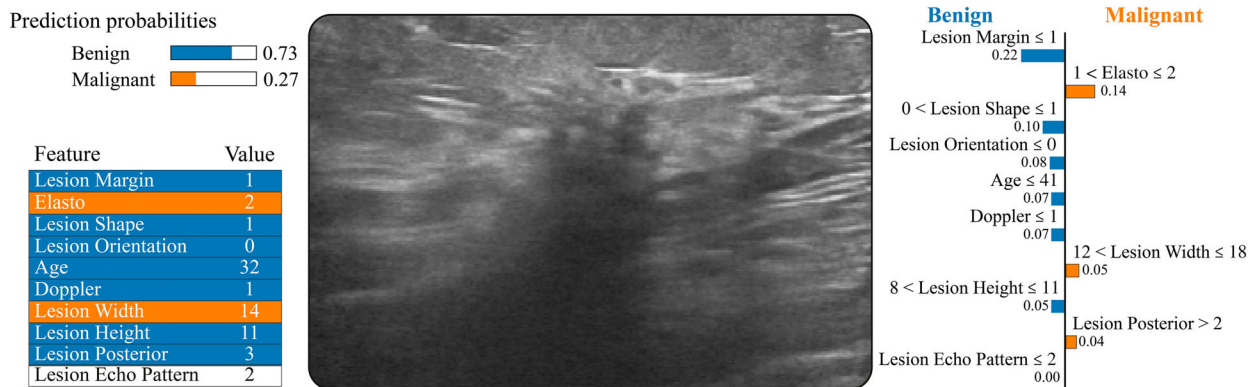
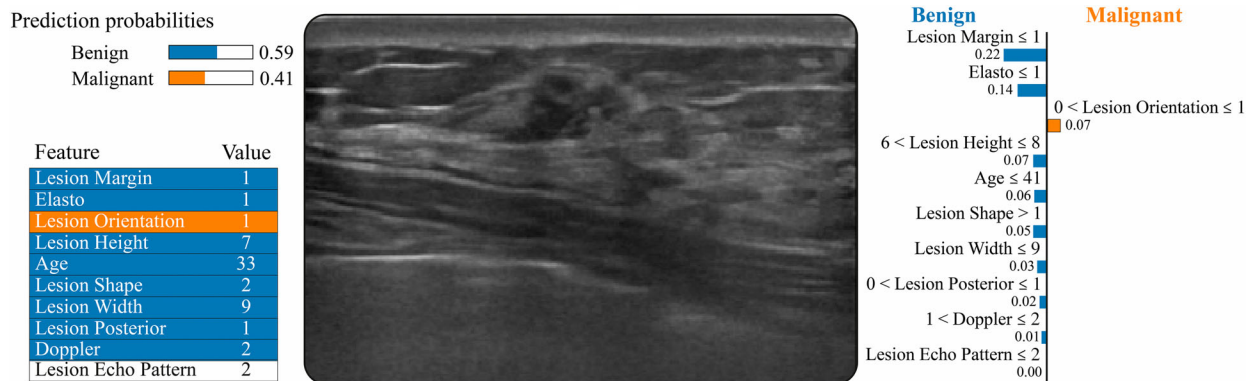


Figure 14. In the ultrasonography image of a 30-year-old woman who presented with breast pain, a hypoechoic, indistinct, not-circumscribed, oval-shaped lesion of 11 × 6 mm in size, with parallel orientation and no posterior feature was observed at 10 o’clock in the left breast. Doppler US showed moderate vascular flow through the lesion. US-Elastography shows that the lesion has a soft structure. The lesion was evaluated as mildly suspicious for malignancy with its contour and vascularization characteristics.



15390613, 2024, 11, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jum.16535 by National Health And Medical Research Council, Wiley Online Library on [07/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

use. To overcome this problem, the features that the classifier with the highest accuracy rate focuses on the test set were analyzed using SHAP. The features that the X²GAI model focuses on in the test set are given in Figure 10. It may be noted that the most focused feature of the X²GAI classifier for the classification of benign and malignant breast lesions is the margin structure of the lesion.

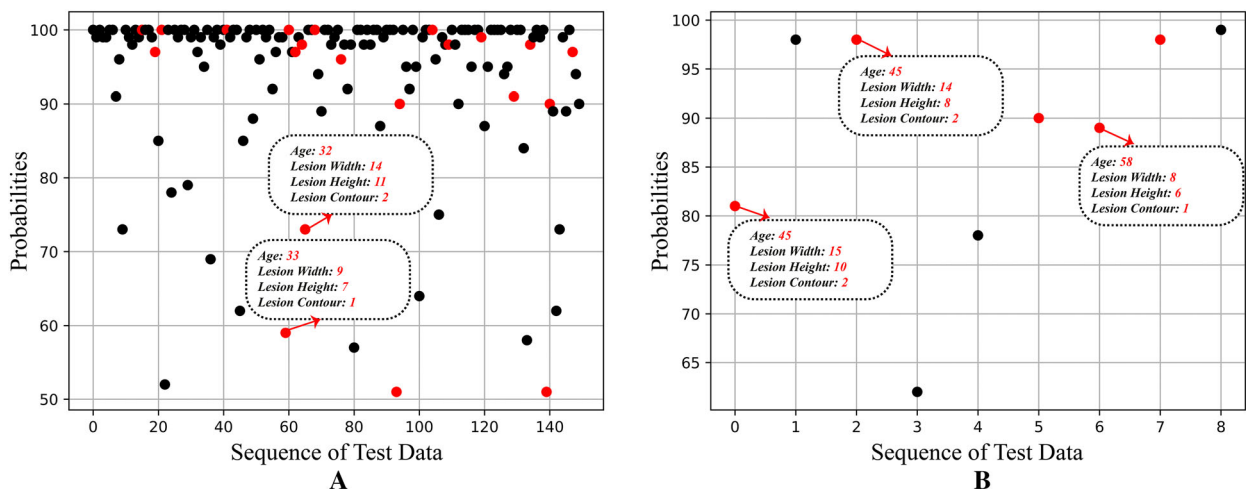
The classifier and radiologists' ranking of the importance of the attributes in the classification process were compared in the study (see Figure 11). Radiologist-1 and Radiologist-2 stated Shape, Margin, and Orientation as the features that should be considered first. X²GAI, on the other hand, considered the "Elasto" attribute in second place, although they agreed on "Margin." Radiologists ranked the "Age" attribute last, while the X²GAI model evaluated this attribute in 4th place. One of the most remarkable results is the "Echo Pattern" attribute, considered important by radiologists and ranked last by XAI.

This difference in priority between radiologists and X²GAI is an important parameter affecting the classification accuracy. X²GAI was able to correctly classify 21 samples that were misclassified by experienced radiologists. Among these cases, ultrasound images and LIME outputs of the model are presented (see Figures 12–14) for cases that were classified as

"malignant" with high probability by the experts but were found to be "benign" by biopsy. In Figure 12, a US-guided lesion biopsy was reported as atypical ductal hyperplasia. Subsequently, total excision was performed and the final pathological result was diagnosed as high-grade ductal carcinoma. Unlike the radiologist, the X²GAI model correctly classified this case correctly (accuracy = 100%). In Figure 13, the lesion was classified as highly suspicious for malignancy due to its radiological features. US-guided biopsy of the lesion was reported as fibrosis. The final pathological result of the lesion, which was excised due to radiological pathological discordance, was diagnosed as a radial scar. The X²GAI model classified this case as benign, with 73%. Figure 13 shows the ultrasound image and the LIME outputs obtained by the proposed model for the sample that was considered malignant by the expert with a lower probability than the other two samples but was found benign after biopsy. Interestingly, it is seen that the X²GAI model also correctly classified this case slightly above the average of 59% but still has difficulty in classification.

It is promising that the X²GAI classifier successfully classified 21 cases where even the experienced expert was wrong. The probability values in the predictions are important to determine the model's

Figure 15. Probabilistic distributions of the predictions made by X²GAI and radiologists on the test data, (A) correctly predicted by X²GAI and incorrectly predicted by radiologist (red circles), (B) incorrectly predicted by both X²GAI and radiologist (red circles). Black circles indicate correct predictions by X²GAI and radiologists.



stability. To evaluate the stability of the X^2 GAI classifier, the prediction probabilities on each test sample were recorded. These probabilities were then divided into two correct and incorrect predictions. The probability values of correct and incorrect predictions of X^2 GAI and radiologists are given in Figure 15.

Discussion

In this article, an explainable breast cancer detection system is developed using a comprehensive breast

cancer dataset (Breast-XD) meticulously collected by radiologists. The classification process performed by the models was made explainable and the findings were compared with the radiologists. Figure 16 compares the classification performances obtained by the main AI algorithms and experienced radiologists using the Breast-XD dataset.

Our proposed X^2 GAI algorithm achieved 94.34% accuracy on test samples randomly selected from the Breast-XD dataset. Moreover, SHAP and LIME methods describe the features they focus on to make an accurate diagnosis.

Figure 16. Performance values of classification results obtained in the study.

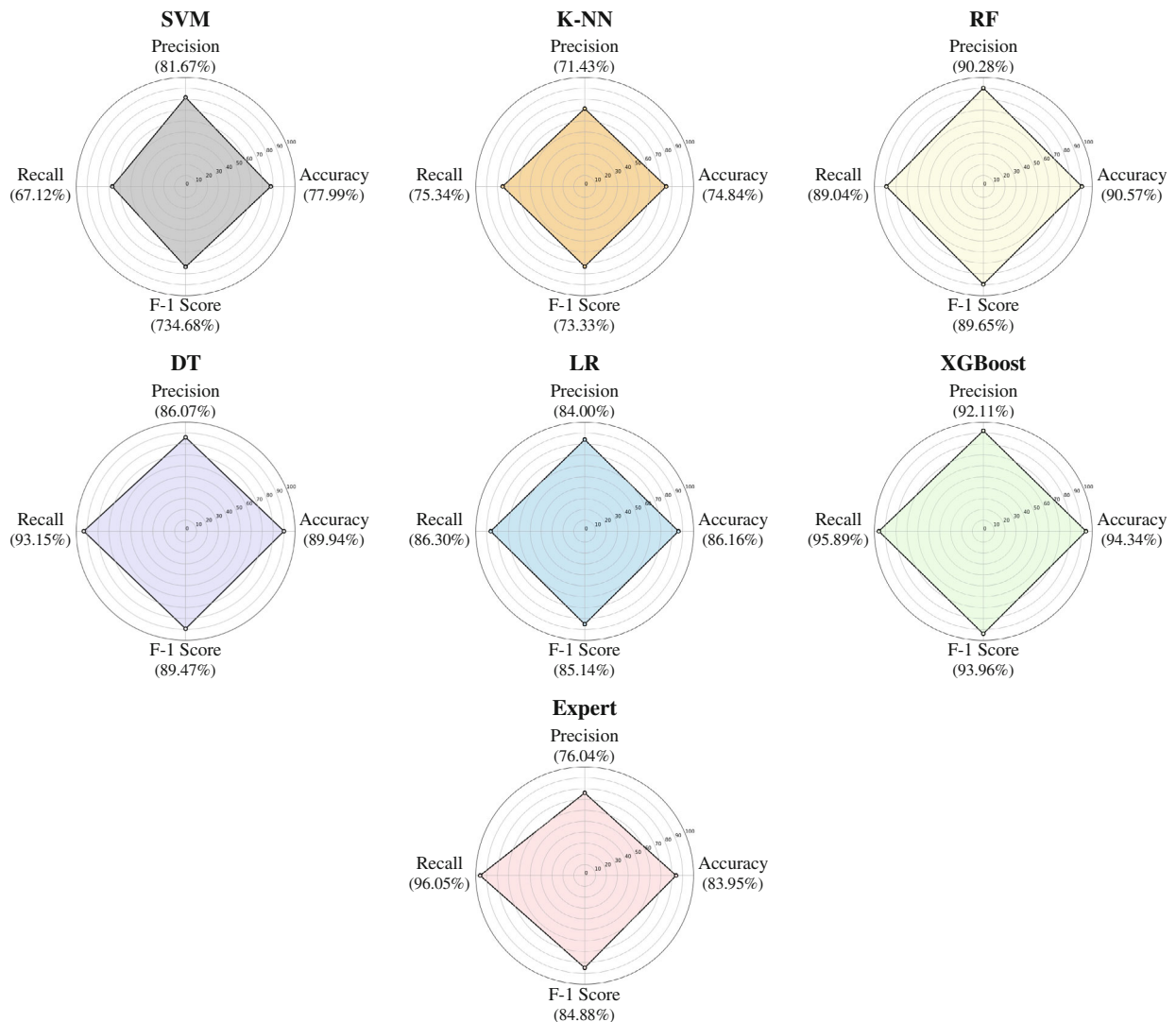
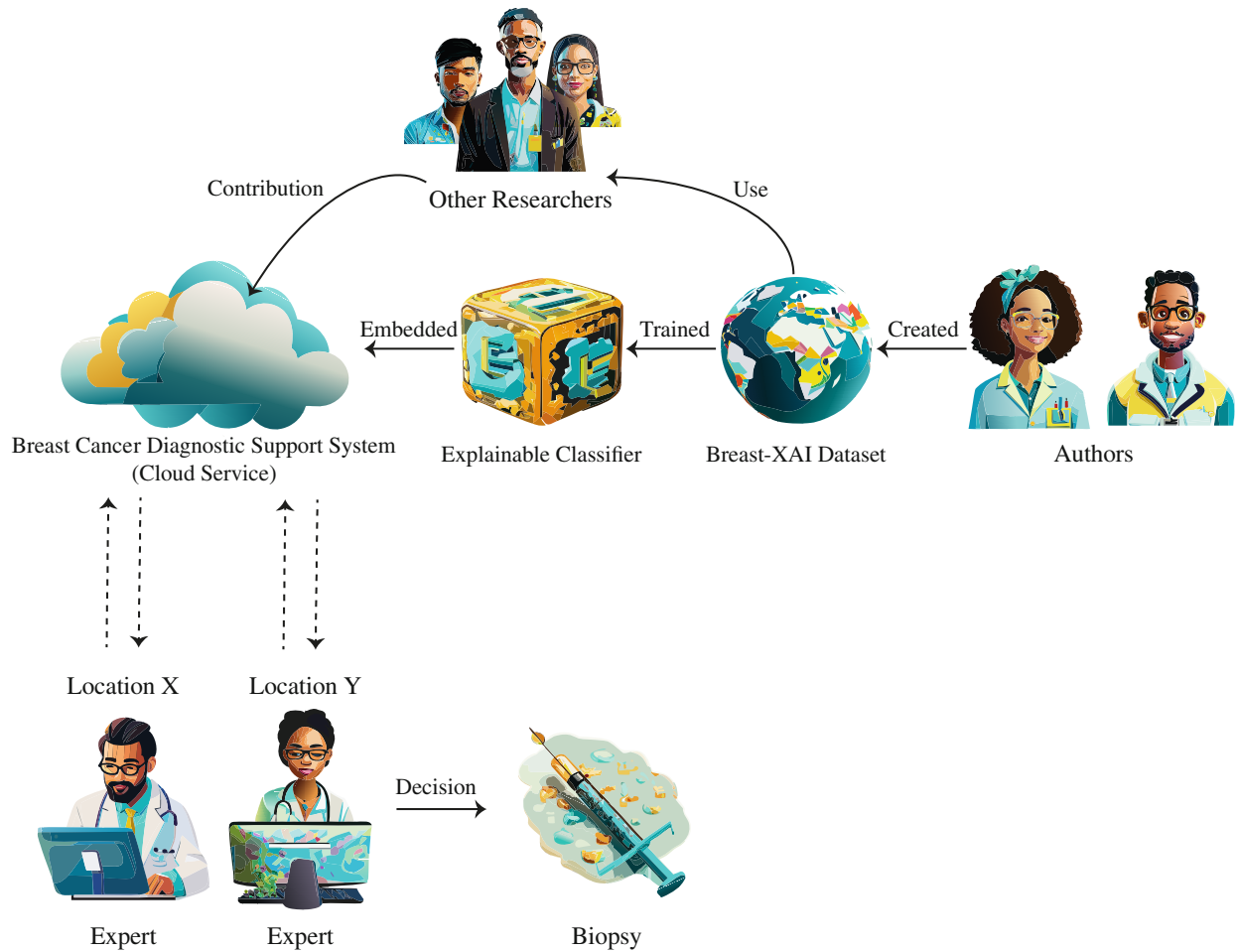


Figure 17. A system structure for the clinical usability of the proposed X^2 GAI model in a cloud environment.



The advantages of the proposed X^2 GAI model for breast cancer detection are summarized as follows:

- It has an automated structure that can obtain more objective results compared with the subjective evaluations of radiologists.
- Compared with radiologists, it provides important information about the importance of the attributes.
- It requires fewer hardware requirements and yields high performance compared with the popular deep learning models.
- Thanks to its explainable structure, it is easy to visualize the features the classifier focuses on during the decision-making phase. In this way, it can

gain the trust of experts during clinical use.

- It can prevent unnecessary biopsies and financial burdens during the classification of benign/malignant lesions.
- The developed model can be used location-independently in the cloud, as shown in Figure 17.

The limitation of the study is that the dataset used in this study was created using a limited set of subjects from one hospital. To obtain more generalized results, it is necessary to create a comprehensive dataset with the participation of different radiologists from different hospitals. In future studies, we intend to create a huge dataset from various centers of different races and age groups.

Conclusion

Breast cancer is a prevalent and lethal type of cancer worldwide. Early detection of cancer is critical as it can save lives by providing timely treatment. In this study, a machine learning-based detection system has been proposed. The results of this study demonstrate that the artificial intelligence-based model developed for the automatic classification (benign/malignant) of breast cancer lesions has significant potential in breast cancer diagnosis. Our proposed ML-based model achieved a high accuracy of 94.34% using a huge private ultrasound dataset. Furthermore, presenting the features on which the model focused during the classification stage in an interpretable structure increased the model's reliability and comprehensibility. This was evidenced by the effective performance of the X²GAI model in cases overlooked by experienced radiologists based on biopsy results. The findings suggest that artificial intelligence can provide significant support to clinicians in diagnosing breast cancer, offering higher accuracy and reliability than traditional diagnostic methods. The developed model has the potential to be employed for real-world clinical applications. Also, the generated method can detect prostate, ovarian, liver, and cervical cancers.

Data Availability Statement

The data that support the findings of this study are openly available in Breast-Cancer at <https://www.kaggle.com/datasets/zalyildirim/breast-cancer-dataset>.

References

- Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2021; 71:209–249.
- Funaro K, Ataya D, Niell B. Understanding the mammography audit. *Radiol Clin* 2021; 59:41–55.
- Hu X, Myers KS, Oluyemi ET, Philip M, Azizi A, Ambinder EB. Presentation and characteristics of breast cancer in young women under age 40. *Breast Cancer Res Treat* 2021; 186:209–217.
- Baş H, Üstüner E, Kula S, Konca C, Demirel S, Elhan AH. Elastography and Doppler may bring a new perspective to TIRADS, altering conventional ultrasonography dominance. *Acad Radiol* 2022; 29:e25–e38.
- Morris EA, Comstock CE, Lee CH, Lehman CD, Ikeda DM, Newstead GM. ACR BI-RADS[®] magnetic resonance imaging. In: Magny SJ, Shikhman R, Keppke AL (eds). *Breast imaging report data system*. American College of Radiology; 2013:5.
- Ji H, Zhu Q, Ma T, et al. Development and validation of a transformer-based CAD model for improving the consistency of BI-RADS category 3–5 nodule classification among radiologists: a multiple center study. *Quant Imaging Med Surg* 2023; 13:3671.
- Manhas J, Gupta RK, Roy PP. A review on automated cancer detection in medical images using machine learning and deep learning based computational techniques: challenges and opportunities. *Arch Comput Methods Eng* 2022; 29:2893–2933.
- Debelee TG, Schwenker F, Ibenthal A, Yohannes D. Survey of deep learning in breast cancer image analysis. *Evol Syst* 2020; 11:143–163.
- Nassif AB, Talib MA, Nasir Q, Afadar Y, Elgendy O. Breast cancer detection using artificial intelligence techniques: a systematic literature review. *Artif Intell Med* 2022; 127:102276.
- Beura S, Majhi B, Dash R, Roy S. Classification of mammogram using two-dimensional discrete orthonormal S-transform for breast cancer detection. *Healthc Technol Lett* 2015; 2:46–51.
- Elmoufidi A, El Fahssi K, Jai-andaloussi S, Sekkaki A, Gwenole Q, Lamard M. Anomaly classification in digital mammography based on multiple-instance learning. *IET Image Process* 2018; 12:320–328.
- Wu WJ, Lin SW, Moon WK. Combining support vector machine with genetic algorithm to classify ultrasound breast tumor images. *Comput Med Imaging Graph* 2012; 36:627–633.
- Huang YL, Wang KL, Chen DR. Diagnosis of breast tumors with ultrasonic texture analysis using support vector machines. *Neural Comput Appl* 2006; 15:164–169.
- Kumar G, Alqahtani H. Deep learning-based cancer detection-relevant developments, trend and challenges. *Comput Model Eng Sci* 2022; 130:1271–1307.
- Yala A, Lehman C, Schuster T, Portnoi T, Barzilay R. A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology* 2019; 292:60–66.
- Abdel-Zaher AM, Eldeib AM. Breast cancer classification using deep belief networks. *Expert Syst Appl* 2016; 46:139–144.
- Ayana G, Dese K, Choe S. Transfer learning in breast cancer diagnoses via ultrasound imaging. *Cancers (Basel)* 2021; 13:738.
- Street WN, Wolberg WH, Mangasarian OL. Nuclear feature extraction for breast tumor diagnosis. In: Acharya RS, Goldof DB (eds). *Biomedical Image Processing and Biomedical Visualization*. Vol 1905. SPIE; 1993:861–870.
- Karabatak M. A new classifier for breast cancer detection based on Naïve Bayesian. *Measurement* 2015; 72:32–36.
- Wang H, Zheng B, Yoon SW, Ko HS. A support vector machine-based ensemble algorithm for breast cancer diagnosis. *Eur J Oper Res* 2018; 267:687–699.

21. Bagui SC, Bagui S, Pal K, Pal NR. Breast cancer detection using rank nearest neighbor classification rules. *Pattern Recogn* 2003; 36: 25–34.
22. Hassija V, Chamola V, Mahapatra A, et al. Interpreting black-box models: a review on explainable artificial intelligence. *Cogn Comput* 2024; 16:45–74.
23. Vollmer S, Mateen BA, Bohner G, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ* 2020; 368:l6927.
24. Saeed W, Omlin C. Explainable AI (XAI): a systematic meta-survey of current challenges and future opportunities. *Knowl Based Syst* 2023; 263:110273.
25. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Adv Neural Inf Proces Syst* 2017; 30:4765–4774.
26. Ribeiro MT, Singh S, Guestrin C. “Why should i trust you?” explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM Digital Library; 2016:1135-1144.
27. Mahesh B. Machine learning algorithms-a review. *Int J Sci Res* 2020; 9:381–386.
28. Cervantes J, Garcia-Lamont F, Rodríguez-Mazahua L, Lopez A. A comprehensive survey on support vector machine classification: applications, challenges and trends. *Neurocomputing* 2020; 408: 189–215.
29. Chandra MA, Bedi SS. Survey on SVM and their application in image classification. *Int J Inf Technol* 2021; 13:1–11.
30. Thomas T, Vijayaraghavan PA, Emmanuel S, Thomas T, Vijayaraghavan PA, Emmanuel S. Applications of decision trees. In: Thomas T, Vijayaraghavan AP, Emmanuel S (eds). *Machine Learning Approaches in Cyber Security Analytics*. Springer; 2020:157-184.
31. Cunningham P, Delany SJ. k-nearest neighbour classifiers—a tutorial. *ACM Comput Surv* 2021; 54:1–25.
32. Boateng EY, Otoo J, Abaye DA. Basic tenets of classification algorithms K-nearest-neighbor, support vector machine, random forest and neural network: a review. *J Data Anal Inf Process* 2020; 8: 341–357.
33. Sharma R, Sharma K, Khanna A. Study of supervised learning and unsupervised learning. *Int J Res Appl Sci Eng Technol* 2020; 8: 588–593.
34. Roy SS, Dey S, Chatterjee S. Autocorrelation aided random forest classifier-based bearing fault detection framework. *IEEE Sensors J* 2020; 20:10792–10800.
35. Amjad M, Ahmad I, Ahmad M, Wróblewski P, Kamiński P, Amjad U. Prediction of pile bearing capacity using XGBoost algorithm: modeling and performance evaluation. *Appl Sci* 2022; 12:2126.
36. Noorunnahar M, Chowdhury AH, Mila FA. A tree based eXtreme gradient boosting (XGBoost) machine learning model to forecast the annual rice production in Bangladesh. *PLoS One* 2023; 18: e0283452.
37. Han Y, Kim J, Enke D. A machine learning trading system for the stock market based on N-period min-max labeling using XGBoost. *Expert Syst Appl* 2023; 211:118581.
38. Holzinger A, Saranti A, Molnar C, Biecek P, Samek W. Explainable AI methods—a brief overview. *International Workshop on Extending Explainable AI beyond Deep Models and Classifiers*. Springer Nature; 2022:13-38.
39. Loh HW, Ooi CP, Seoni S, Barua PD, Molinari F, Acharya UR. Application of explainable artificial intelligence for healthcare: a systematic review of the last decade (2011–2022). *Comput Methods Prog Biomed* 2022; 226:107161.
40. Pelegrina GD, Duarte LT, Grabisch M. A k-additive Choquet integral-based approach to approximate the SHAP values for local interpretability in machine learning. *Artif Intell* 2023; 325:104014.
41. Li Z. Extracting spatial effects from machine learning model using local interpretation method: an example of SHAP and XGBoost. *Comput Environ Urban Syst* 2022; 96:101845.
42. Zafar MR, Khan N. Deterministic local interpretable model-agnostic explanations for stable explainability. *Mach Learn Knowl Extr* 2021; 3:525–541.