scientific reports

Check for updates

OPEN

A Deep Learning and Explainable Artificial Intelligence based Scheme for Breast Cancer Detection

Sandeep Saharan[1]✉, Niyaz Ahmad Wani[1], Shreeya Chatterji[1], Neeraj Kumar[1] & Abdullah Mohammed Almuhaideb[2]

Incorporating Artificial Intelligence (AI) presents significant potential for transforming multiple aspects of the healthcare sector, encompassing administration, medical prediction, decision-making, and diagnostics. However, its often perceived enigmatic nature is a significant barrier to the widespread use of AI in the medical field. AI systems, specifically deep learning models, have demonstrated exceptional performance, often comparable to that of humans. However, their internal mechanisms sometimes resemble opaque entities, leading to challenges in understanding and placing trust in their decision-making processes. The presence of skepticism has limited their practical application in medical contexts. Given the challenges above, a pioneering proposed scheme, namely "*D*XAIB", has been introduced. The proposed scheme utilizes a hybrid methodology that integrates Convolutional Neural Networks (CNNs) with the Random Forest (RF) model to detect breast cancer accurately. One distinguishing characteristic of "*D*XAIB" is its distinctive focus on interpretability of the predictions. The RF technique, employed to predict class labels, implements an automated feature learning process facilitated by the "*D*XAIB" scheme, which has several convolutional layers. However, "*D*XAIB" doesn't stop at providing accurate predictions. To tackle the significant concern surrounding the interpretability of AI, an approach called "SHAP" is utilized to offer explanations at both the local and global levels for every prediction. This implies that "*D*XAIB" not only facilitates the identification of a diagnosis but also provides comprehensive explanations regarding the rationale behind a specific prediction, augmenting transparency and fostering confidence in the decision-making process of the AI system. The proposed "*D*XAIB" scheme surpasses other cutting-edge schemes in terms of the prediction outcomes it achieves and, therefore, in terms of its explainability findings.

An intelligent healthcare system is defined by integrating sophisticated technology, including data analytics, the Internet of Things (IoT), and Artificial Intelligence (AI), with conventional medical procedures. Implementing this shift improves patient outcomes by facilitating predictive diagnosis, personalized treatment, and real-time patient monitoring. Advancements in intelligent technology enhance communication and optimize healthcare processes, resulting in financial savings and heightened effectiveness. Ultimately, this transition prompts healthcare professionals to adopt a more proactive approach and depend on data analytics. Hence, the intelligent healthcare system enables the prompt detection of illnesses, thereby reducing their severity and transmission patterns. Therefore, it assumes a pivotal function in the mitigation of epidemics and the safeguarding of public health. Furthermore, an efficiently operating healthcare system improves efficiency by mobilizing individuals with improved medical conditions, promoting economic progress and social growth. Scientific and medical breakthroughs have led to vast medical revolutions, improving treatment quality and decreasing costs. Recent advancements in oncology have offered promising prospects and efficacious therapies for serious illnesses such as cancer. Undoubtedly, cancer is well recognized as the second most common cause of death. According to recent predictions from the US National Cancer Institute, the anticipated number of newly diagnosed cancer cases in 2024 is around 2,001,140, resulting in an expected 611,720 deaths[1]. Given its high mortality rates, breast cancer is the second leading cause of cancer-related deaths in women[2]. Breast cancer is characterized by the uncontrolled proliferation of cells in the breast tissue, exhibiting several types classified based on the specific abnormalities seen in the cells. Approximately 90 percent of instances are associated with age-related genetic anomalies, whereas hereditary factors generate the remaining 5 to 10 percent. Diagnosing a disease depends

[1]Department of Computer Science and Engineering, Thapar Institute of Engineering and Technology, Patiala, Punjab 147004, India. [2]Department of Networks and Communications, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia. ✉email: sandeepsaharan@outlook.com

on comprehensively examining clinical findings and carefully implementing experimental procedures. About 42,250 people will lose their lives to breast cancer in 2024, with an expected 310,720 new cases, according to the American Cancer Society[3]. Effective identification and forecasting of medical conditions are essential for optimizing cost control and developing comprehensive treatment strategies. The use of automated diagnostic technologies might resolve the primary challenges related to the identification and categorization of breast cancer. Despite the considerable body of research on the identification of breast cancer, further work is required to develop more effective and practical approaches. Implementing these techniques would enhance decision-making, leading to favorable patient and family outcomes. Consequently, specialists from various disciplines, including medicine and science, have demonstrated significant interest in the challenges associated with the detection of breast cancer. Analyzing correlations among datasets enables Machine Learning (ML) systems to tackle complex issues effectively. Advanced Deep Learning (DL) models need significant data and high-performance workstations[4]. As long as these criteria are met, DL can identify or categorize numerous medical diseases[5–7], including alzheimer[8], breast cancer[9,10]. So far, many ML and DL algorithms have been used to forecast breast cancer in women. Naive Bayes (NB), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Gradient Boosting (GB), and Convolutional Neural Networks (CNNs) are often used algorithms in this framework. In the recent past, breast cancer prediction systems have been developed using both image and tabular datasets. Torabi et al.[11] designed the Self-Supervised Adversarial Adaptation Network (SelfAdaptNet) as a specialized model for identifying breast cancer cases in mammography pictures. The SelfAdaptNet model integrates Semi-Supervised Learning (SSL) approaches to tackle the challenge of acquiring large quantities of precisely annotated information in medicine. Additionally, it employs adversarial learning to effectively transfer information from a large, unlabeled source domain to a separate target domain, enhancing performance across varying domain distributions. Arshad et al. assess deep learning models for breast cancer diagnosis by analyzing five CNN architectures to determine the most precise one. DenseNet-121 attained the greatest accuracy at 99%, followed by VGG-16 at 98%, underscoring the potential of artificial intelligence to enhance diagnostic precision and assist pathologists in resource-constrained environments such as Pakistan. The methodology introduced by Qian et al.[12] employs confocal electron microscopy (CEM) images to predict the status of biomarkers, including HER2, PR, ER, and Ki-67 in breast cancer. This approach was evaluated using a dataset consisting of 1203 samples. This strategy performed better than the approach that employed a CNN network without the suggested limitation of lesion attention. In the latter approach, the inputs were derived from a bounding box that included the locations of CEM lesions. Mahmood et al.[13,14] tackles the diagnostic issues of breast cancer by using deep learning algorithms, such as the Chaotic Leader Selective Filler Swarm Optimization and hybrid neural network methods, to enhance lesion identification and classification in mammograms.

The method known as Stacking Ensemble Meta-Learning (SEMeL-LR), first proposed by Prusty et al.[15], blends the expected results of basic ML models for the best possible outcomes. This methodology has five clearly defined stages. The classification of breast cancer included six well-established models: Liver Response (LR), Support Vector Machine (SVM), Neural Network (NB), K-Nearest Neighbors (k-NN), CART, and a recently developed model named SEMeL-LR. The researchers used the open-source Kaggle online platform to get their hands on the original Wisconsin Breast Cancer Dataset (WBCD). In their groundbreaking work, Iqra et al.[16] presented MOB-CBAM, a deep learning architecture incorporating two attention mechanisms. The implemented architectural approach has shown exceptional effectiveness in accurately forecasting breast cancer. The model exhibits the ability to provide accurate predictions with a significant degree of precision, therefore enabling it to effectively identify many types of breast cancer, including masses, calcifications, and both benign and malignant diseased growths. In order to determine which ML algorithms were most effective in correctly classifying breast cancer tumors, Mohi Uddin et al.[17] used the WBCD. To evaluate the performance of the models, many measures were created. A quantitative study was undertaken on many algorithms to determine the most efficient methodology. This vote classifier achieves the utmost level of accuracy. The prediction of survival rates in persons diagnosed with breast cancer has been thoroughly examined in previous research[18,19]. In their study, Manzoet al. (2023) proposed a survival model that utilizes ML to forecast the probability of cancer patients surviving. The comprehensive multi-modal fused system presented by Yuan et al. (2023) serves as a prediction model for evaluating the overall survival rate after five years of breast cancer patients. The study undertaken by Gyorffy et al.[20] employed treatment-based groups of patients to ascertain and rank all prognostic biomarkers and genes that might assist in forecasting the future survival of individuals with breast cancer. Mahmood et al.[21] developed a ConvNet-based deep learning model combined with SVM to improve the precision of breast cancer diagnosis in mammography, achieving 97.7% training accuracy and 97.8% validation accuracy. This technique outperforms previous models such as VGGNet and ResNet, assisting pathologists in minimizing diagnostic mistakes and avoiding needless biopsies.

Although DL outperforms ML in many areas, both are considered "Black Boxes" because they cannot explain their prediction accuracy[22]. Consequently, healthcare practitioners possess a restricted comprehension of the output generated by the model. Therefore, medical practitioners, especially doctors, who depend on evidence-based diagnoses to guide treatment choices, are reluctant to adopt AI models. Consequently, it is crucial to explain the outcomes produced by models to build trust and encourage using these systems in many fields. It is all the more critical when a single action has the potential to endanger lives. ML in healthcare raises issues about fairness, confidentiality, privacy, and comprehensibility. Considering the significant impact of these systems' decision-making processepredictionealth, it is crucial to comprehend them[23]. Consequently, transparency and explainability have become crucial elements of AI in the healthcare sector. AI systems in healthcare must exhibit precise forecasts and offer explicit reasons for each prediction since explainability is of paramount significance[24]. This work integrates DL with eXplainable Artificial Intelligence (XAI) techniques to identify breast cancer patients precisely and provide meaningful explanations for the predictive diagnosis. Implementing this scheme enhances medical professionals' comprehension and practical application of the scheme's outcomes.

Motivation

While efficient, conventional diagnostic techniques (for example mammography, breast ultrasound, breast biopsy, breast Magnetic Resonance Imaging (MRI), blood tests, CT scan) may be time-consuming and lack the accuracy required for tailored therapy. ML algorithms can analyze extensive data and detect nuanced patterns, making them a potent tool for more precise and early prediction of breast cancer. In intelligent healthcare systems, many systems already predict breast cancer efficiently. Further, medical practitioners use prediction systems for early diagnosis, which greatly decreases the cost of diagnosis as very few cases will go for further conventional diagnosis. There exist many systems where the accuracy of predicting breast cancer is already very high. However, in such critical applications, we always look for the development of a more accurate, robust system that produces no false pessimistic predictions. Thus, recall becomes an essential factor to measure the performance of such systems. False pessimistic predictions are the ones that can endanger life as after looking at the predictions, the patients/ doctor may or may not go for the conventional diagnosis, which can deteriorate the stage of cancer. Nowadays, the body blindly believes in the genuineness of the predictions, and everyone, including medical practitioners and patients, looks for the reasons and explanations behind such predictions. The importance of explanations increases further if the cancer prediction is optimistic. Thus, before making a conventional diagnosis of breast cancer, medical practitioners and patients greatly need reasons/ explanations. Also, explanations can reduce the effect of false pessimistic predictions to a great extent, as medical practitioners can find the correctness of such predictions by looking at their explanations. They promote transparency, enhance confidence among healthcare practitioners, and facilitate well-informed decision-making while treating breast cancer patients. The comparison of notable works in the area of breast cancer detection is provided in Table 1. Based on this comparison two primary research questions were framed and focused while doing present work. These are: How hybrid and ensemble of machine or deep learning models can effectively predict the malignant cases of breast cancer with better recall/ sensitivity rate? How XAI methods can assist physicians/ patients for better understanding behind produced predictions?

Contributions

The main contributions of the presented work are as follows:

1. Employing a versatile hybrid model, known as "*D*XAIB", which combines the CNN framework with the RF technique, substantial enhancements in the accuracy and efficiency of breast cancer detection have been achieved. This novel methodology integrates the benefits of Convolutional Neural Networks (CNN) and Random Forests (RF) by using CNN's expertise in feature extraction and RF's robust ensemble learning protocols.
2. The hybrid approach integrates the widely used XAI technology, "SHAP," providing healthcare practitioners with comprehensive explanations and valuable insights. The use of SHAP enables the generation of easily understandable and explainable forecasts. This feature enhances the ability of physicians to make informed clinical choices by comprehending the many aspects that impact the outcomes.
3. This study used an XAI methodology systematically implemented to assess the "Breast Cancer Wisconsin (Diagnostic)" dataset. The "*D*XAIB" hybrid model demonstrates superior prediction effectiveness compared to conventional state-of-the-art approaches, signifying a significant advancement in breast cancer detection.

System Model and Problem Formulation

The system model of this research work contains three layers: a data layer, a prediction layer, and an explainability layer, as shown in Fig. 1. These layers are explained as follows:

1. Data layer: This layer provides breast cancer data for time-to-time training and testing. Also, this layer will record and deliver the vitals of new patients to the next layer, i.e., the prediction layer. This layer may also record the outcome of the new patient whose vitals it provided to the prediction layer.

Proposals ↓ Parameters →	1	2	3	4	5	6	7	8	9	10
Torabi et. al.[11]	✓	✗	✓	✗	✓	✓	✗	✗	✗	✗
Mahmood et. al.[14]	✓	✗	✓	✗	✓	✓	✓	✓	✗	✗
Yuan et. al.[19]	✓	✗	✓	✓	✓	✓	✓	✓	✗	✗
Singh et. al.[25]	✓	✓	✗	✓	✓	✓	✓	✓	✗	✗
Manikandan et. al.[26]	✓	✓	✗	✗	✓	✓	✓	✓	✗	✗
Qian et. al.[12]	✓	✓	✓	✗	✓	✗	✗	✗	✓	✗
Mahmood et. al.[13]	✓	✗	✓	✓	✓	✓	✓	✓	✓	✗
Nissar et. al.[16]	✓	✗	✓	✗	✓	✓	✓	✓	✓	✗
Mahmood et. al.[21]	✓	✗	✓	✓	✓	✓	✓	✓	✓	✗
*D*XAIB: The proposed scheme	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1. Comparison with existing proposals. 1: Machine Learning, 2: Ensemble Learning, 3: Deep Learning, 4: Hybrid Learning, 5: Accuracy, 6: Recall, 7: Precision, 8: F1 score, 9: Explainability, 9: SHAP, ✓: considered, and ✗: non-considered

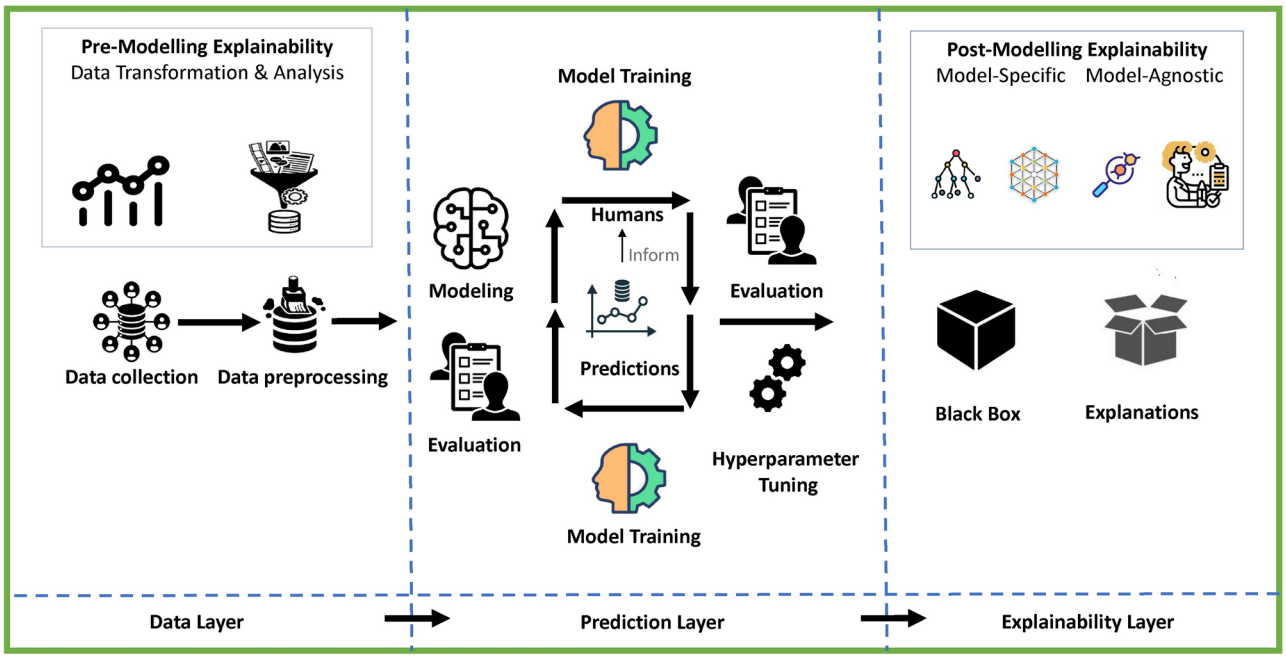

Fig. 1. *D*XAIB: System Model.

2. Prediction layer: This layer will train *D*XAIB model at regular time intervals using the available dataset. After testing of *D*XAIB model, it will be used to get breast cancer status given the various vitals of the patient. This model can used by both patients and medical practitioners for the outcome of breast cancer status given the multiple vitals of the patient.
3. Explanability layer: Explainability in ML models refers to the ability to understand and interpret the decisions/ predicction made by models, particularly in complex or critical domains. This layer is essential for building trust, ensuring accountability, and enabling stakeholders to comprehend how a model arrives at its predictions or classifications. This layer will allow patients and medical practitioners to logical reasoning behind the prediction outcome from the *D*XAIB model. It will typically explain the role of all individual input vitals in the prediction outcome of the *D*XAIB model.

Let us consider the dataset is denoted by 'D' with '$\mathtt{m}$' rows and '$\mathtt{n}$' columns where $(\mathtt{m},\mathtt{n}) = \mathtt{size}(\mathrm{D})$. Let d_{ij} is data instance present at i^{th} row and j^{th} column of the dataset '$\mathtt{D}$'. Let the vitals be given in columns of the dataset '$\mathtt{D}$'. Whereas the values of these vitals for all individual patients are given in the rows of the dataset '$\mathtt{D}$'. The vitals considered in this work are '$\mathtt{radius}$', '$\mathtt{texture}$', '$\mathtt{perimeter}$', '$\mathtt{area}$', '$\mathtt{smoothness}$', '$\mathtt{compactness}$', '$\mathtt{concavity}$', '$\mathtt{concave\ points}$', '$\mathtt{symmetry}$', and '$\mathtt{fractal\ dimension}$'. These all parameters are related to the breast. The outcome can be the type of breast cancer, '$\mathtt{benign}$' or '$\mathtt{malignant}$'. After splitting the data into training and testing data, let both types of data be denoted by using '$\mathrm{D_{tr}}$' and '$\mathrm{D_{ts}}$' respectively. Let us consider the numbers of data instances in the training and testing dataset are m_{tr} and m_{ts} respectively. Hence, $m = m_{tr} + m_{ts}$. Let 'l_i' is the likelihood of class 'o', i.e. '$\mathtt{benign}$', and '$(1 - l_i)$' is the likelihood of class '$(1 - o)$', i.e., '$\mathtt{malignant}$' for the 'i^{th}' data instance. Let ∇_L denote the log loss or binary cross entropy. Then, the objective function of this work is given as follows in Eq. (1).

$$\text{Minimize } \nabla_L = \frac{1}{m_k} \sum_{i=1}^{m_k} -(o_i \times \log(l_i) + (1 - o_i) \times \log(1 - l_i)), \tag{1}$$

subject to constraints:

$$0 \leq l_i \leq 1, \tag{2}$$

$$o_i \in \{0, 1\}, \tag{3}$$

$$0 < m_k \leq m_{tr} \tag{4}$$

The constraint given in Eq. (2) defines that the probability or likelihood of any class should be between '$\mathtt{0}$' and '$\mathtt{1}$'. The constraint given in Eq. (3) defines that the outcome in the present case of research can be either '$\mathtt{0}$' or '$\mathtt{1}$' where the former is the notation for the benign and the later is for malignant breast cancer. Whereas, the last constraint given in Eq. (4) defines the calculation of binary cross entropy on a number of instances, which can be, at most, the number of instances that are there in training data. This number depends on the model's training in every epoch.

*D*XAIB: The Proposed Scheme

This section provides a detailed description of the *D*XAIB scheme, including the utilized data-set, its pre-processing, the forecasting of breast cancer, and the explainability of the prediction made by the proposed scheme.

Data-set Description

This study used a dataset entitled "Breast Cancer Wisconsin (Diagnostic)" obtained from the UCI ML repository[27]. The dataset contains a comprehensive collection of 30 clearly defined features, including the mean, standard error, and worst values of 10 specific features (`texture`, `radius`, `concavity`, `area`, `smoothness`, `compactness`, `symmetry`, `perimeter`, `concave points`, and `fractal dimension`). These features are obtained from ultrasound pictures of a breast mass obtained by fine needle aspiration (FNA). These features delineate the cellular nuclei seen in the photographs. The collection has a total of 569 unique individual samples. The comprehensive dataset provided a solid basis to follow-up analysis since each item comprised several relevant features. The class label functioned as a categorical variable, indicating separate and identifiable groupings. The class label consisted of two well-defined categories, each representing distinct diagnostic results. The category denoted as 'M'refers to all presently surviving females who have received a diagnosis of breast cancer. Individuals categorized as 'B' indicate the lack of breast cancer. The Breast Cancer Wisconsin (BCW) dataset was chosen for several key reasons such as, historical significance and validation of it, its simplicity and interpretability, and it is a clean and well-structured data that make it particularly useful for the development of classification models for breast cancer detection. The BCW dataset has been extensively used and validated in breast cancer research. The dataset contains a comprehensive set of features derived from fine needle aspiration (FNA) of breast masses, focusing on cell characteristics such as radius, texture, and perimeter, which are clinically relevant for identifying malignancies. The dataset's features are interpretable, making it ideal for initial testing and development of ML models. The BCW dataset is highly structured, with no significant data inconsistencies, which allows for straightforward analysis and model development.

Data Pre-processing

Data preprocessing encompasses many procedures, including cleaning, transformation, and preparation. The aforementioned steps are carried out to ensure the appropriate processing and structuring of the raw data for further analysis. Through the assurance of data consistency and logical progression, this essential process enables rigorous model training and precise predictions. Approaches such as normalization, feature engineering, and managing missing values are employed to improve the usability and efficiency of datasets for training ML models. A scaling methodology was applied to the dataset following rigorous data collection and a thorough feature selection approach. This measure was implemented to maintain consistency and facilitate relevant comparisons of the characteristics. Scaling, or standardization, was used to consistently measure all traits and minimize biases caused by differences in measuring units. The scaling method included subtracting the mean value from each feature and then dividing the resulting value by the standard deviation of the feature. The methodology employed in this work successfully maintained the relative relationships among the various characteristics and conserved the overall organization of the dataset. Statistical scaling techniques provide an equitable and impartial evaluation of the correlations and patterns seen in the dataset. By standardizing all characteristics to a consistent scale, the potential impact of variations in measurement units was reduced, enhancing the study's trustworthiness. Quantitative representations of categorical data in '`D`' are obtained via label encoding. Specifically, the target column in '`D`' comprises categorical values 'M' and 'B' corresponding to malignant and benign outcomes. The application of label encoding results in assigning the value '1' to the category 'M' and the value '0' to the category 'B' (algorithm 1: line 3). Label encoding converts categorical data into a format appropriate for various ML approaches. An additional pre-processing stage involves removing superfluous features from the dataset '`D`' (algorithm 1: lines 4). Following label encoding and feature selection, all occurrences of duplicate data are eliminated from the dataset '`D`'. In present type of critical applications the sensitivity or recall plays a very critical role where no positive case should be predicted as negative case which can eventually lead to fatality later. The used dataset in the present work is having 62.85% of benign cases whereas, 37.15% of malignant cases. So, dataset was biased towards benign class. Thus, Synthetic Minority Over-sampling Technique (SMOTE) which is a data augmentation technique for tabular data-set is used. It created synthetic samples for minority malignant class by interpolating between existing data points. It can enhance recall by generating synthetic examples for the minority class in an imbalanced dataset. Instead of simply duplicating existing instances, SMOTE creates new, synthetic data points by interpolating between nearest neighbors of the minority class. This process expands the feature space of the minority class, helping models learn more diverse patterns and reducing the chance of misclassifying minority instances as negatives.

Input : D, N, R, c_{loss}, $c_{optimizer}$, r_{lr}, $c_{metrics}$, $c_{monitor}$, r_{factor}, $n_{patience}$, $n_{verbose}$, n_{epochs}, n_{batch_size}, n_estimators, max_depth, min_samples_split
Output : rf_predictions, accuracy, recall, precision, F1_score, shap.force_plot, shap.summary_plot

```
Load the dataset D
(m,n) = size(D)
Replace (Malignant, Benign) with (1,0) in dataset D respectively        #Label Encoding
D = D/D.id                                                              # Delete id from Dataset D
X = D/D.diagnosis
y = D.diagnosis
X_train, y_train, X_test, y_test = Split_Dataset(X, y, test_size = 0.2)
n_class = 10                                                            # Define the number of classes
cnn_model = Sequential([                                                # Creation of CNN Model
      Add Conv1D Layer with Input Shape
      Add Conv1D Layers
      Add MaxPooling1D Layers
      Add Dropout Ratio
      Add Dense Layers
      Add Dense Output Layer with number of classes equals to n_class
])
cnn_model.compile(c_loss, c_optimizer, r_lr, c_metrics)                 # CNN Compilation
reduce_lr = ReduceLROnPlateau(c_monitor, r_factor, n_patience, n_verbose)
cnn_history = cnn_model.fit(X_train, to_categorical(y_train, n_class), n_epochs, n_batch_size, validation_data=(X_test,
to_categorical(y_test, n_class)), callbacks = [reduce_lr])
cnn_features_train = cnn_model.predict(X_train)
cnn_features_test = cnn_model.predict(X_test)
param_grid = {
     Set some values for parameters, such as, 'n_estimators', 'max_depth', 'min_samples_split'
}
grid_search = GridSearchCV(RandomForestClassifier(), param_grid=param_grid, cv=3)
grid_search.fit(cnn_features_train, y_train)
best_rf_model = grid_search.best_estimator_
rf_predictions = best_rf_model.predict(cnn_features_test)
accuracy = accuracy_score(y_test, rf_predictions)
recall = recall_score(y_test, rf_predictions)
precision = precision_score(y_test, rf_predictions)
F1_score = f1_score(y_test, rf_predictions)
import shap
shap.initjs()
import SelectKBest
selector = SelectKBest(k=10)                                            # Select the top 10 features
selector.fit(X_train, y_train)
X_train_reduced = selector.transform(X_train)
explainer = shap.Explainer(best_rf_model, cnn_features_train)
shap_values = explainer(cnn_features_test)
import SelectKBest
selector = SelectKBest(k=10)                                            # Select the top 10 features
selected_instance_repeated = np.repeat(selected_instance, 114, axis=0)
selector.fit(selected_instance_repeated, y_test)
selected_instance_reduced = selector.transform(selected_instance_repeated)
shap_values[0]
shap.initjs()
explainer_hybrid = shap.TreeExplainer(best_rf_model)
shap_values = explainer_hybrid.shap_values(X_test)
import shap
shap.initjs()
shap.force_plot(explainer_hybrid.expected_value[1], shap_values[1][5], X_test.iloc[5, :])
SHAP_values_hybrid = explainer_hybrid.shap_values(X_test)
shap.initjs()
shap.force_plot(explainer_hybrid.expected_value[0], SHAP_values_hybrid[0], X_test)
shap.summary_plot(SHAP_values_hybrid[0], X_test, plot_type='dot')
```

Algorithm 1. *D*XAIB: The Proposed Scheme.

Breast Cancer Detection

A hybrid ML model that combines Convolutional Neural Networks (CNNs) for feature extraction from tabular dataset with Random Forest for predictive output is developed in this present work. To enhance transparency and interpretability, SHAP (SHapley Additive exPlanations) explainable AI module is included which offers comprehensive insights into the determinants influencing the model's prediction results. The proposed scheme starts with the loading of data, which is then followed by pre-processing steps. The dataset, denoted as 'D', has 'm' rows and 'n' columns, with 'm' indicating the number of rows or data instances and 'n' indicating the number of

columns or features. Following the successful importation of 'D', pre-processing steps are executed as explained in the preceding sub-heading. Subsequently, the data set 'D' is partitioned into two subsets: one containing the target feature, including the target class label, and the other including the other features. The proposed *D*XAIB scheme partitions the dataset into two subsets for training and testing purposes. This is accomplished by using an 80-20 split (algorithm 1, line 7). After the split, data is normalized using the MinMaxScaler function. Following data transformation, the dataset 'D' is partitioned into four subsets: X_train, y_train, X_test, and y_test (algorithm 1, line 7). The X_train and y_train datasets are fed into the proposed CNN model, which is a part of the *D*XAIB scheme. Various layers in the CNN architecture, including convolutional, flattened, dense, and max-pooling layers, include dropout rates. The architectural design of the proposed CNN can be seen in Table 3. Importantly, the proposed CNN is only used for feature learning rather than data categorization. A total of 10 output classes are included in the proposed CNN model. Next, the classification layer of the CNN model is substituted by the RF classifier. The CNN under consideration consists of a total of 10 layers, which include four convolutional layers, one flattened layer, three dense layers, and two max-pooling layers. However, an excessively high number of layers may harm performance, so balancing enhancing accuracy and allocating resources is essential. To address the issue of over-fitting, a dropout rate of '0.20' is applied following every max-pooling and dense layer. After passing through the convolutional and max-pooling layers, the input transforms a planar structure before being sent to the first thick or dense layer. Three thick layers are present, with the first two levels consisting of '512' and '256' number of filters, respectively. After the proposed CNN architecture is built, the model is trained using appropriate training data. The validation technique, as shown in algorithm 1, is conducted using validation data with a batch size of '64' epochs and '100' epochs. Before transmitting the data to the RF classification layer, a dense layer is used to modify and restructure it after the training phase. An essential function of this layer is to synchronize the results of the CNN model with the necessary vector shape for the RF layer (algorithm 1: lines 9). Several factors justify the selection of the RF model as the classifier layer, namely its superior accuracy when applied to the dataset compared to other ML techniques. Subsequently, the classification layer generates projections for the resultant class, and the precision of these projections, which measure the model's adaptive performance, is evaluated. Indeed, the suggested approach is an essential antecedent and decisive factor in directing future evaluations of breast tumor recipients. These patients receive a range of crucial clinical procedures specifically intended to accomplish a thorough assessment. Importantly, our first goal is to determine whether a patient has cancer. Having formulated predictions, this work focuses on clarifying the importance of specific features in identifying patients with various types of cancer.

Explainability

After the class prediction process is completed, the findings or projections are transmitted to the eXplainable Artificial Intelligence (XAI) algorithm for the proposed scheme. Subsequently, the XAI framework is used to provide explanations. The SHAP (SHapley Additive exPlanations) XAI technique is used in this work over others because of its theoretical foundation based on Shapley values, its consistency and additivity, its global and local interpretability, model-agnostic and model-specific variants, its visualization and interpretability, its reliability on feature importance, its ability to capture feature interactions, and its wide adoption and continuous improvement[28]. It is used in the proposed scheme to engage with the RF layer and provide various explanations. The SHAP approach offers local and global reasons, as shown by the statistical data in the findings section. While the patient may find these explanations unclear, they are crucial for medical practitioners. Upon analysis by a healthcare professional, these clarifications allow them to offer the patient many justifications for a specific disease diagnosis and suggest several alternative diagnostic possibilities (algorithm 1: lines 24-47). After the identification or forecasting stage of the proposed method is finished, the SHAP approach is used to analyze and present many explanations for the predictions produced by the model. SHAP, a well acknowledged and effective method in Comprehensible Artificial Intelligence , is especially developed to uncover the precise influence of certain characteristics on the results produced by a model. The objective of installing SHAP was to achieve accurate and easily understandable explanations in order to improve the transparency of the decision-making framework in the existing system. Prior to implementing SHAP, the proposed ML model predicted outcomes for all patients in the dataset. After that, we calculated SHAP values for all of the model features to see how much of an effect each piece of information had on our prediction technique. The SHAP values measure the effect of each feature's presence or absence on the model's output, allowing us to evaluate their relevance. Furthermore, the SHAP scores were meticulously categorized into interpretations for the identified results. More specifically, throughout the stage of some patients, interpretations are generated, offering a comprehensive examination of the main characteristics of each prognosis. Through offering comprehensive explanations customized for each patient, physicians may ascertain the fundamental factors that impacted the diagnosis of each individual patient. By taking this route, they were able to better comprehend how the model arrived at its conclusions. In order to uncover the most prominent trends and patterns, the SHAP values were averaged throughout the sample. Through this aggregate, explanations were extracted on a global scale, revealing underlying features that are regularly linked to forecasts of breast cancer type, whether it is malignant or benign. These global explanations aimed to emphasize overarching patterns in the data and enhance understanding of the system's functioning in a broader scope. Incorporating SHAP reasoning after the proposed system's classification phase has shown to be a great help in giving a comprehensive and comprehensible assessment of the scheme's predictions. The SHAP technique provided clinicians and researchers with profound insights at the individual patient level and on a broader scale. This enabled them to enhance their trust in the model's evaluations and develop a more profound comprehension of the complex connections between traits and pathological outcomes. Transparency and comprehensibility are crucial for establishing confidence and, hence, determining the practical feasibility and capability of the system to improve breast cancer detection and patient care.

Parameters	Values
c_{loss}	'categorical_crossentropy'
$c_{optimizer}$	Adam
r_{lr}	0.0001
$c_{metrics}$	'accuracy'
$c_{monitor}$	'val_loss'
r_{factor}	0.1
$n_{patience}$	5
$n_{verbose}$	1
n_{epochs}	100
n_{batch_size}	64
n_estimators	[100, 110, 120, ... 480, 490, 500]
max_depth	[None, 5, 6, 7 ... 18, 19, 20]
min_samples_split	[2, 4, 6, ... 16, 18, 20]

Table 2. Learning parameters and values.

Layer	Kernel size	Filter count
Conv1D, Conv1D, Conv1D, Conv1D	3 × 3	128, 128, 256, 256
MaxPooling1D, MaxPooling1D	2 × 2	–, –
Dense1, Dense2, Dense3	–	512, 256, 2

Table 3. Architecture of CNN embedded in *D*XAIB.

Performance Evaluation

This part provides a comprehensive performance assessment of the proposed method, including its system design, experimental settings, diagnostic data, and explanations.

System Configuration

The value of various hyper-parameters used in learning process of proposed scheme are given in Table 2. It is important to mention here that logistic chaotic maps is used for hyper-parameter tuning process.

Experimental Parameters

The architectural design presented in this study comprises 4 Conv1D layers, two MaxPooling1D levels, and 3 Dense layers, as seen in Table 3 of the reference. The suggested technique consists of two initial convolutional layers that are fitted using '128' filters. The next two convolutional layers consist of convolutional filters with a size of '256'. Following each pair of convolutional layers, the max-pooling layer is integrated. Within this particular max-pooling architecture, the pooling size is always set at a constant value of (2,2). The model's convolutional layers are activated using the Rectified Linear Unit (ReLU) function and feature a kernel size of (3,3). A dropout rate of 0.20 has been involved in the design.

Before being fed into the first thick layer, the data is flattened and dropout is applied. The next dense layers include filters of sizes '512', '256', and '2', respectively. The first two dense layers use the Rectified Linear Unit (ReLU) activation mechanism. Whereas, the final dense layer uses the 'Softmax' one. During the model compilation process, Adam serves as the optimizer and uses a category-based cross-entropy loss estimator.

State-of-the-art Schemes

Comparisons were made between the outcomes of the proposed scheme and state-of-the-art techniques utilizing many assessment criteria, including accuracy, precision, recall, and F1-score. It can be seen in Table 4. The following section presents the state-of-the-art schemes. The results of the suggested system were compared against state-of-the-art models using many evaluation metrics, including accuracy, precision, recall, and F1-score. Refer to Table 4 for visualization.

1. Yuan et al.[19]: This research suggests using a deep multi-modal fusion network (DMMFN) to assess breast cancer patients' five-year survival rate. Medical data, copy number variation data, and gene expression data are used to make the prediction. The imbalanced dataset issue is solved via SMOTE-NC oversampling in the first step. In order to extract abstract modal features from multi-modal input, a 2-layer, one-dimensional CNN and a bi-directional LSTM network are used. The last stage of the MaxoutMLP classifier includes fusion characteristics for accurate prediction.
2. Setiono[29]: This study introduces a neural network pruning method that reduces network complexity while improving breast cancer diagnosis accuracy. The shortened network rules are accurate like the original

Model	Accuracy	Precision	Recall	F1 score
Naive Bayes	0.9212	0.9214	0.9316	0.9321
Random Forest	0.9416	0.9212	0.9323	0.9323
Support Vector Machine	0.8321	0.8423	0.8765	0.8862
Light Gradient Boosting Machine	0.9234	0.9444	0.9567	0.9435
Yuan et al.[19]	0.9369	0.8913	0.8454	0.8677
Setiono[29]	0.9500	–	–	–
Hasan et al.[30]	0.9795	–	–	–
Aalaei et al.[31]	0.9730	–	–	–
Naik et al.[32]	0.9350	–	–	–
Singh et al.[25]	0.9831	0.9800	–	0.9539
Kiyan et al.[33]	0.9760	–	–	–
Karabatak et al.[34]	0.9740	–	–	–
Gibbons et al.[35]	0.9600	0.9400	0.9900	–
Obaid et al.[36]	0.9810	–	–	–
Hasan et al.[37]	0.9068	0.9240	0.9690	0.9460
Wani et al.[38]	0.9829	0.9872	0.9872	0.9872
Manikandan et al.[26]	0.9800	0.9900	0.9700	0.9700
Chaurasia et al.[39]	0.9516	–	–	–
Liu et al.[40]	0.9580	–	–	–
Abdar et al.[41]	0.9807	0.9810	–	0.9810
Sahu et al.[42]	0.9700	–	–	–
Nilashi et al.[43]	0.9410	–	–	–
Sakri et al.[44]	0.8130	0.8830	–	0.8760
Ramadevi et al.[45]	0.9789	–	–	–
Murugesan et al.[46]	0.9776	0.9881	–	0.9800
Idris et al.[47]	0.9453	0.9429	–	0.9413
Rajaguru et al.[48]	0.9561	0.9726	–	–
Proposed Scheme (*D*XAIB)	0.9835	0.9876	0.9874	0.9872

Table 4. Performance evaluation of contemporary methods.

network. The simplified criteria achieve above 95% diagnosis accuracy on breast cancer training and test datasets.

3. Hasan et al.[30]: This study introduces an ensemble classification method. The ensemble technique from principal component analysis uses the Rotation Forest algorithm. This technique uses multilayer perceptron neural networks instead of decision trees as foundation classifiers, unlike rotation forest topologies. In this arrangement, PCA creates different feature sets from the original dataset, enhancing the ensemble classifier system.
4. Aalaei et al.[31]: This study focusses on features selection for breast cancer detection. It wraps a PS-classifier with a Genetic Algorithm (GA) for feature selection. The experimental results reveal that this model performs similarly to other models on Wisconsin breast cancer datasets.
5. Naik et al.[32]: Wrapper-based fitness functions for feature selection are presented in the study. This function differs from most existing methods that prioritize accuracy alone, which may fail with imbalanced datasets. The sensitivity and specificity entropy are integrated instead. This method seeks to equilibrate True Negative and True Positive Rates. The proposed fitness function was assessed using four single-pass classifiers. Feature selection consistently enhanced performance.
6. Singh et al.[25]: This work introduces a novel feature selection method that combines two algorithms and six ML classifiers. This method improves efficiency and comprehensiveness of selected attributes. The three algorithms use their information to remove superfluous traits.
7. Kiyan et al.[33]: This study uses the Wisconsin breast cancer dataset to evaluate the radial basis function, general regression neural network, and probabilistic neural network topologies. This dataset is popular in ML, neural networks, and signal processing. These statistical neural networks aim to increase breast cancer diagnosis accuracy and impartiality.
8. Karabatak et al.[34]: This study introduces an automated breast cancer diagnosis method using Association Rules (AR) and Neural Networks (NN). AR reduces breast cancer dataset dimensions, whereas NN facilitates intelligent categorization. Comparing the AR with NN system against a single NN model yields a classification accuracy of 95.6%. The results show that AR reduces feature count and that AR plus NN models may be utilized to construct efficient and automated diagnostic systems for various ailments.
9. Gibbons et al.[35]: The authors utilized a publicly available dataset with few inputs and cases. This dataset helps medical practitioners compare statistical and ML methods. The dataset's small size ensures quick computational processing on most present processors.

10. Obaid et al.[36]: This study tested Support Vector Machine, K-nearest Neighbors, and Decision Tree for breast cancer classification. The Wisconsin Breast Cancer (Diagnostic) dataset was used. The goal was to determine the most accurate classifier. Results show that the Quadratic Support Vector Machine achieved the best accuracy of 98.1% and lowest false discovery rates.
11. Hasan et al.[37]: The authors mined breast cancer incidence data from the Surveillance, Epidemiology, and End Results database post 2004. This work developed a Multilayer Perceptron (MLP) ML model to reliably predict breast cancer survival.
12. Wani et al.[38]: The authors created a hybrid DL breast cancer detection method. This study's hybrid strategy made more accurate diagnoses. The authors used SHAP to improve prediction model explainability.
13. Manikandan et al.[26]: An effective ML breast cancer classification approach is presented in this work. Additionally, this research employed two-step feature selection. The study identified the most important breast cancer dataset features using Variance Threshold and Principal Component Analysis. The breast cancer dataset is classified using supervised and ensemble learning methods after feature selection. These approaches include ADA Boosting, XGBoost, Gradient Boosting, Naive Bayes, and Decision Tree. ML algorithms are evaluated using train-test split and k-fold cross-validation.
14. Chaurasia et al.[39]: This study evaluates six ML algorithms (CART, SVM, NB, KNN, LR, and MLP) on the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. Classification test accuracy, standardized data correctness, and runtime performance are evaluated. This work aims to increase prediction accuracy by applying a novel statistical feature selection method.
15. Liu et al.[40]: A new breast cancer diagnosis technique uses an Information Gain-directed Simulated Annealing Genetic Algorithm Wrapper (IGSAGAW) to choose the most important characteristics. The method uses the Information Gain algorithm to choose the best 'm' features. Cost-Sensitive Support Vector Machine (CSSVM) learning algorithms leverage these characteristics.
16. Abdar et al.[41]: ML and data mining are used to automate breast cancer prediction in this study. To differentiate benign and malignant breast tumors, the researchers employed layered ensemble approaches using Stacking and Voting classifiers.
17. Sahu et al.[42]: A prediction model using AI and multivariate statistics is proposed in this work. Given the noise in datasets from diverse sources, data mining is crucial to automate the diagnostic process. A unique hybrid feature selection strategy combining PCA and ANN is presented in the study. Data is preprocessed and the most important characteristics are identified using PCA. The Wisconsin Breast Cancer Dataset from the UCI Machine Learning Repository was used for classification with 10-fold cross-validation. The approach outperformed numerous classifiers in accuracy, sensitivity, and F-measure, making it more effective for diagnosis.
18. Nilashi et al.[43]: This work proposes a knowledge-based breast cancer classification technique that combines clustering, noise reduction, and classification methods. The Expectation Maximization (EM) method groups data by similarity. CART is used to build fuzzy rules. A fuzzy rule-based reasoning system classifies breast cancer using these principles.
19. Sakri et al.[44]: A number of well-known data mining technologies are being tested to predict breast cancer recurrence. To increase prediction model accuracy, the system blends particle swarm optimization for feature selection into three popular classifiers: naive Bayes, K-nearest neighbors, and fast decision tree learner.
20. Ramadevi et al.[45]: In this work, authors tested PCA, a popular feature extraction method, on breast cancer datasets. To assess PCA's influence on breast cancer data, several classifiers are used and the results analyzed.
21. Murugesan et al.[46]: In this work, authors created a CAD system using a super learner to detect illness. The experiment used seven UCI ML datasets: WDBC, SHD, HCC, HD, VCD, CHD, and ILP. Each dataset was preprocessed and split into training and testing sets. The authors used a wrapper-based feature selection technique with three bioinspired algorithms-CSO, KH, and BFO-and the SVM classifier's accuracy to find the optimum feature subsets.
22. Idris et al.[47]: The ID3 approach struggles to detect continuous-valued data and improve decision tree classification accuracy. This work seeks to fix this. FID3 explicitly uses the ID3 algorithm for decision tree learning and utilizes fuzzy logic. The research evaluates the FID3 algorithm using WBCD (Original), WDBC (Diagnostic), and Coimbra breast cancer datasets. By compared to other methods, the FID3 algorithm is effective and high-performing. FID3 and FUZZYDBD offer a reliable and robust approach that categorizes breast cancer well.
23. Rajaguru et al.[48]: This study seeks to construct a potent breast cancer classification ML model. The paper addresses classification problems using tree-based and non-parametric ML. PCA for feature selection on the UCI Wisconsin Breast Cancer Diagnosis dataset improves dataset variance and binary classification, as shown in the study. The proposed model uses decision trees and KNN algorithms to demonstrate that the slow-learning KNN method outperforms the decision tree approach.

Evaluation Metrics

The metrics used in this work to evaluate the performance of the proposed scheme with state-of-the-art schemes are as follows:

1. Accuracy: It quantifies the ratio of accurately identified cases to the total occurrences in the dataset. It offers a comprehensive assessment of model efficacy, especially beneficial when classes are equilibrated.
2. Precision: It quantifies the precision of positive predictions, indicating the ratio of real positive forecasts to the total cases classed as positive. Precision is especially advantageous in scenarios where false positives incur significant costs.

Fig. 2. SHAP force plot for class '1' generated using *D*XAIB.

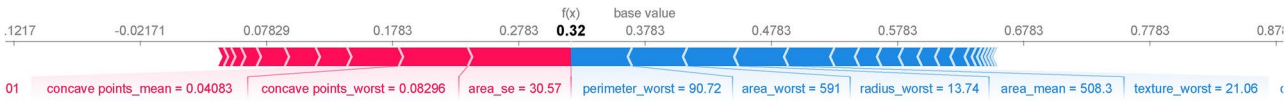

Fig. 3. SHAP force plot for class '0' generated using *D*XAIB.

3. Recall: It measures the model's capacity to detect all relevant positive events within the dataset. It is the ratio of accurate positive forecasts to all actual positive cases, making it significant in situations when false negatives incur substantial costs.
4. F1 Score: It is the harmonic mean of accuracy and memory, offering a singular metric that equilibrates both. It is particularly advantageous in scenarios with unequal class distribution or where both false positives and false negatives incur significant costs. The F1 Score ranges from 0 to 1, with higher values indicating an improved equilibrium between precision and recall.

Diagnosis Results

This sub-section thoroughly explains the outcomes acquired using the proposed scheme, *D*XAIB.The results are presented comprehensively, including numerical summaries for detailed study. The *D*XAIB scheme exhibits exceptional performance in several metrics, such as '0.9835' accuracy, '0.9876' precision, '0.9874' recall, and '0.9872' F1 score. The current study demonstrates the proposed scheme's resilience and usefulness in properly diagnosing breast cancer cases. The review is enhanced by Table 4, which offers a comparison analysis demonstrating the substantial performance superiority of the *D*XAIB system relative to more established systems in the field. The evaluation employs many benchmarks to guarantee a comprehensive assessment of the proposed scheme's effectiveness. The assessed approach underwent a training and testing phase of about '100' epochs. Moreover, the study employs rigorous training and testing protocols, guaranteeing exceptional accuracy and uniformity. Furthermore, this research evaluates other ML models, and the *D*XAIB scheme demonstrates superior performance to those ML models across many evaluation metrics, as seen in Table 4. It is essential to mention that the proposed scheme's achieved recall value decreased in false negative cases. The more excellent value of such cases can create panic among patients and medical practitioners and, thus, can lead to increased cost of further conventional diagnosis. The K-fold cross-validation technique enhances the reliability and validity of the findings, hence strengthening the suggested framework. These convincing results prove the suggested breast cancer detection technique is reliable and effective. The *D*XAIB system outperforms existing schemes in improving diagnostic accuracy and healthcare outcomes.

Explainability Results

Local and global explainability studies exist. The next subsections will clarify these groups.

Local Explanations

Local interpretability involves explaining the unique qualities and reasoning behind a given data instance, not recognizing overall trends or patterns. This study enriches existing knowledge by providing profound insights into the determinants that result in particular outcomes or predictions in individual instances, crucial for managing complex data analysis. As seen in Figs. 2 and 3, discrete force charts provide local explanations for patient decision-making within a dataset. The graphical representations reveal SHAP (SHapley Additive exPlanations) values and the various factors that predict or categorize a patient's condition. SHAP values help understand systems by revealing the factors that impact predictions or classifications. The characteristics in question serve as the decisive criteria for categorizing patients into either class '0' or class '1'. The latter class represents the desired result or trend of interest. Figures 2 and 3 depict two separate situations that provide convincing evidence for categorizing particular patients into specific categories. After rigorous data analysis, it is clear that the prediction is affected by certain attributes, indicated by red, which favors class '1' outcome. Conversely, the forecast is biased towards class '0' because of characteristics linked to the color blue. The features shown in the figures are the ones whose impact is more on the outcome of the prediction model.

Additionally, it is imperative to mention that Figs. 2 and 3 explicitly define the model's fundamental baseline value, which is '0.3783' in the present case. This value acts as threshold for classification. Category `1' is designated for situations in which the anticipated or predicted value exceeds `0.3783'. Selected instances with values equal to or below '0.3783' are classified as `0'. This thresholding technique significantly affects the conversion of model forecasts into categorical class representations. Further elucidation is provided by the force charts seen in Figs. 2 and 3, highlighting the features that have the most influence or significance in determining a patient's categorization. The provided information is essential for understanding the model's decision-making

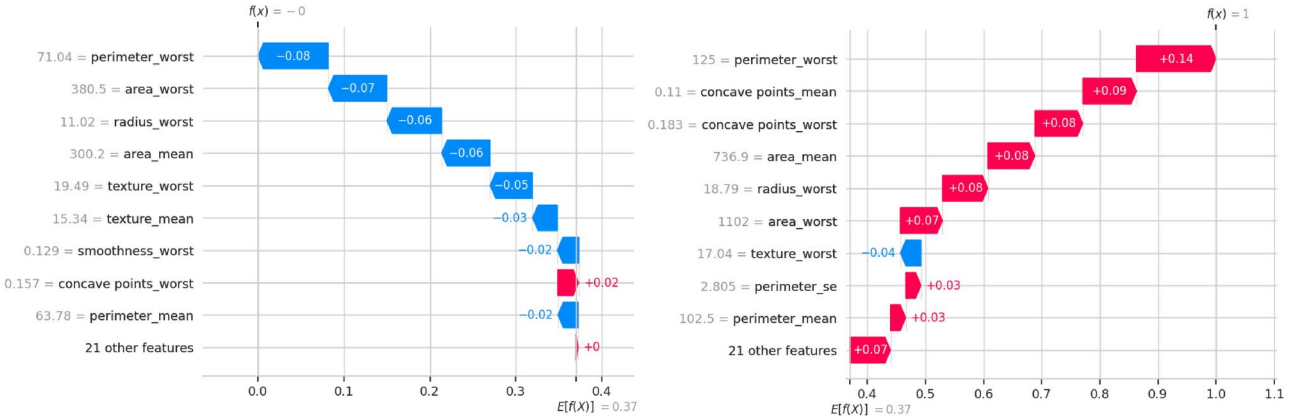

Fig. 4. Waterfall force plot for two instances generated using *D*XAIB.

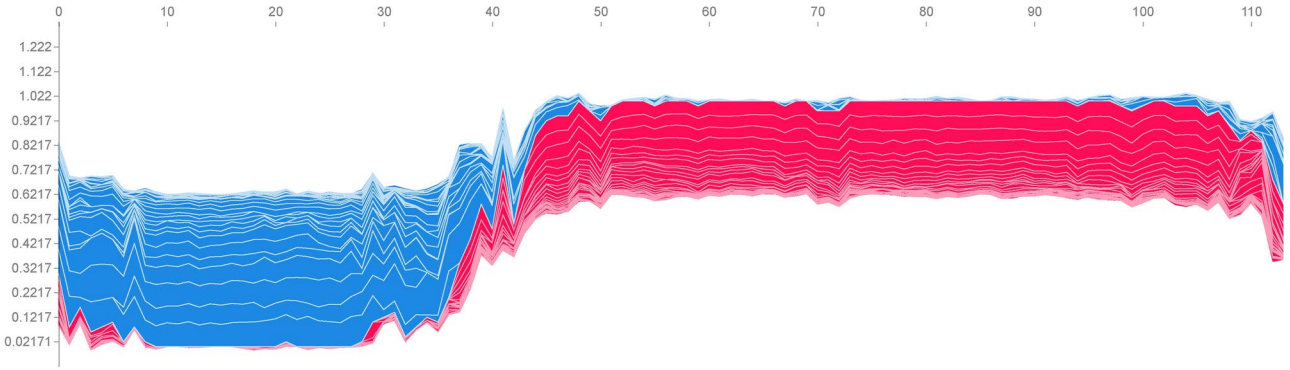

Fig. 5. Global similarity plot generated using *D*XAIB.

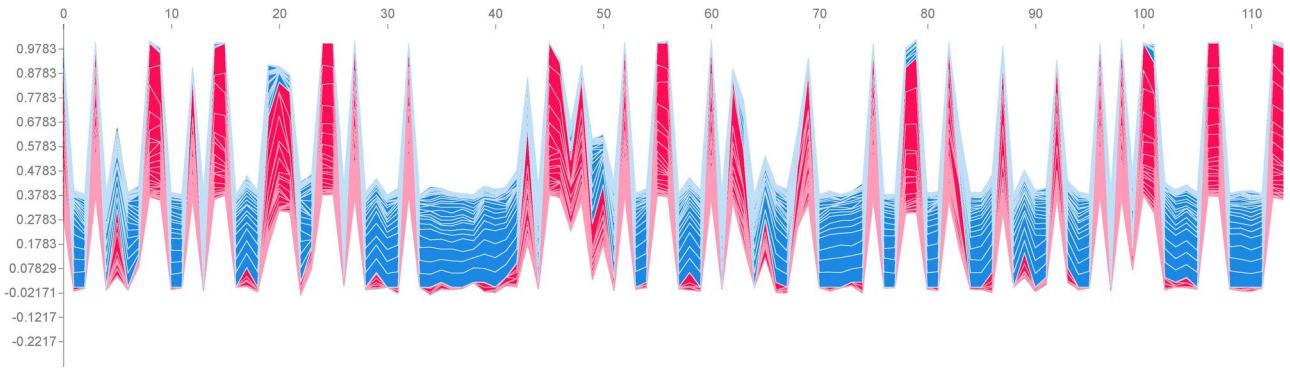

Fig. 6. Global ordering plot generated using *D*XAIB.

process at a specialized and customized level. Enhanced openness provides a valuable understanding of the factors that impact specific forecasts.

The study's Fig. 4 concentrates on a specific group of patients and investigates the impact of important features on the scheme's predictions. The statistical representations in this image elucidate the difference in SHAP values, thereby substantiating the significance of each feature. With respect to the impact of numerous features, these comprehensive explanatory pictures provide a detailed representation of the scheme's predictions and the anticipated outcomes. The waterfall graph, depicted in Fig. 4, categorizes features based on the value of their SHAP ratings, with the lower section primarily containing scores that are more miniature. The features that contribute to the prognosis of breast cancer are represented by red color and in contrast, the blue color represents features that are associated with the negative outcome of breast cancer. These illustrations offer an exhaustive

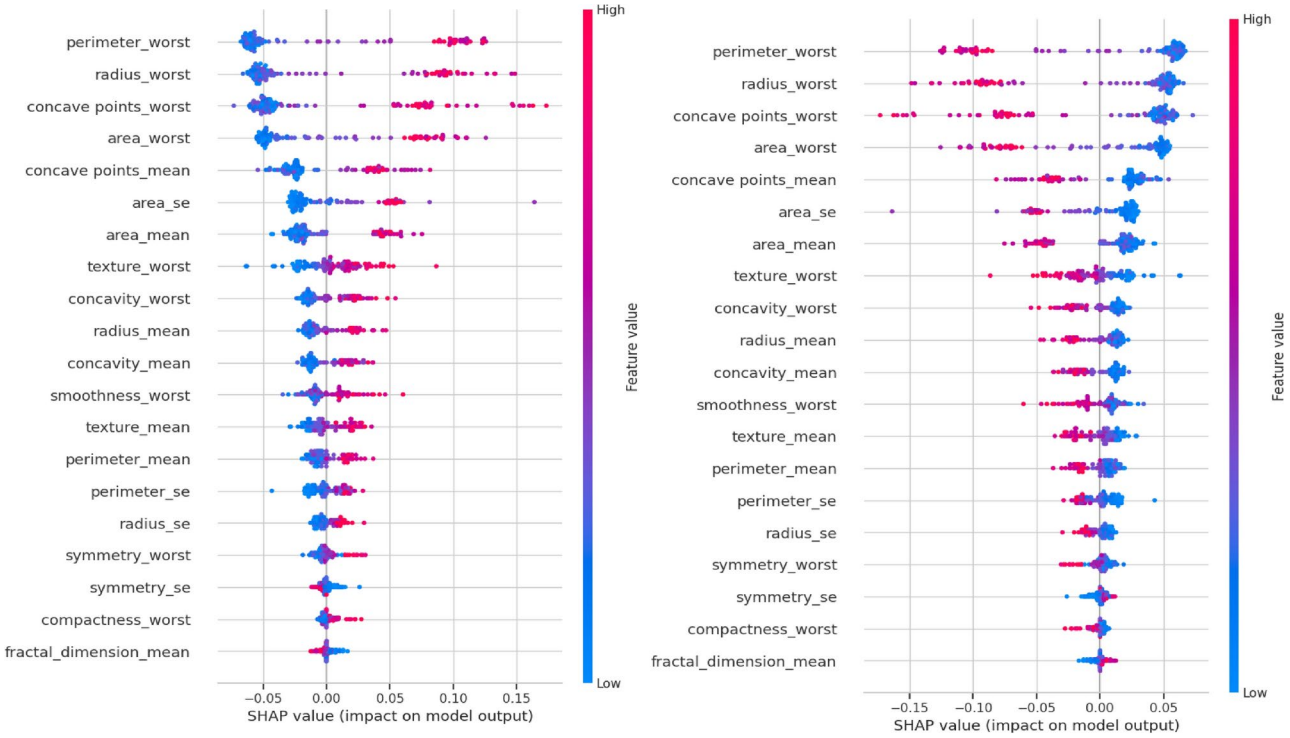

Fig. 7. Global summary plot generated using *D*XAIB.

representation of the scheme's predictions and the extent to which specific features influence them, thereby enhancing the ability to understand and influence medical decisions.

Global Explanations

Figures 5 and 6 demonstrate the actual use of global interpretability utilizing SHAP. According to the analogy of their explanations, the graphic above depicts the grouping of 115 occurrences. Any patient whose cumulative value is above the threshold of '0.3783' is categorised as class '1'. In contrast, those whose aggregate score falls below this level are categorized as class '0'. Figure 5 visualizes the grouping of patients according to the resemblances in their forecasts, therefore facilitating a vital understanding of the patterns in the data. The graph illustrates that patients with values between '1' and '40' often prefer clustering. This phenomenon is mostly attributed to the frequent occurrence of the blue color within this particular spectrum. In contrast, instances that fall beyond these specified boundaries are classified as a distinct category, mostly because of the higher frequency of the red color. The evidence shown in Fig. 6 demonstrates a consecutive organization of samples, spanning from '0' to '115', which unveils consistent trends in the distribution of explanations across the dataset. The proposed work enhances the existing knowledge of explainable machine-learning mechanisms. The thorough comprehension of these intricate interconnections offers useful insights into the possible application of global interpretability in healthcare decision-making. The present results emphasize the urgent need for clear and understandable ML models in medical environments, to enhance patient results and foster confidence in healthcare solutions implemented by artificial intelligence.

Figure 7 offers a comprehensive depiction of the specific number of characteristics that influence the predictions generated by the suggested method. An exhaustive summary of the ongoing debate on the factors that affect the model results may be found in Fig. 7. To demonstrate each characteristic's varying levels of impact, this picture juxtaposes variables of lower rank with those of higher rank. The deliberate use of color coding improves the clarity of the graph. Within this scheme, the color blue represents lesser values of characteristics, whereas the color red signifies greater values. Furthermore, blue dots indicate negative SHAP values, whereas red dots emphasize good ones. Moreover, the degree of horizontal dispersion of data points in the picture provides valuable insights into the relative significance of each factor in shaping the model's predictions. A wider variance of data points indicates a greater impact, while values that are tightly clustered around zero indicate less significant characteristics. This comprehensive presentation successfully elucidates the intricate relationships among numerous characteristics and their influence on the predicted outcomes of the model. SHAP visualizations were used to augment radiologists' comprehension of model predictions including many factors pertinent to breast cancer diagnosis. Utilizing SHAP, we can assess the impact of each feature on the model's choice for specific predictions, providing radiologists with a transparent understanding of how certain qualities affected the model's confidence in categorizing patients as benign or malignant. This feature-specific analysis assists radiologists in identifying the main factors influencing the model's rationale, allowing them to juxtapose these AI-generated insights with conventional diagnostic standards. Furthermore, SHAP visualizations aid in

detecting possible biases or anomalous feature significance in the model's behavior, hence assisting radiologists in enhancing diagnostic procedures with an added degree of transparency. This concordance between model predictions and clinical comprehension eventually cultivates confidence in AI-assisted diagnosis and assists radiologists in making better informed, precise clinical judgments.

Conclusion

ML alone cannot meet healthcare prediction requirements. Validating interpretability is crucial for full understanding of model processes and eliminating mistakes or biases in outputs. XAI is an approach that enhances the transparency and clarity of AI systems. This proves useful in medical diagnostics, improving end-user understanding. This study introduces an interpretable DL technique for classifying breast cancer in females. There are two main goals of this work. The first goal is to improve predicted accuracy, precision, and recall measures. The second goal is to ensure the correctness and comprehensibility of the predictions for individual patients and community at large. A thorough system evaluation indicated remarkable performance in meeting breast cancer diagnostic requirements. This scheme combines a CNN with a supervised RF classifier to acquire features and recognize classes. Additionally, the proposed scheme incorporates interpretability using the model-agnostic SHAP technique. The present scheme recognizes many factors that lead to breast tumor formation. This scheme aims to improve the early planning phase, which is crucial for coordinating healthcare, allocating resources, and providing help to those with diseases. This scheme may improve the quality of clinical care and treatment effectiveness for patients. The employed dataset lacks diversity due to limited imaging modality, demographic and clinical heterogeneity, and non-real-time data obtained at a particular moment in time. Actual breast cancer diagnosis frequently entails longitudinal research, analyzing data from various time periods (e.g., tumor development). The current study has limitations that will be addressed in future breast cancer diagnostic research.

Data availibility

The code generated in the present work is with the corresponding author and may be supplied upon a reasonable request. The breast cancer and other datasets used in this investigation are appropriately referenced in the included references section.

Received: 5 September 2024; Accepted: 19 November 2024
Published online: 01 September 2025

References

1. Institute, N. C. Seer stat fact sheets: All cancer sites. https://seer.cancer.gov/statfacts/html/all.html (2024).
2. WHO. Who breast cancer data. https://www.who.int/news-room/fact-sheets/detail/breast-cancer (2024).
3. Institute, N. C. Seer stat fact sheets: Breast cancer stats. https://seer.cancer.gov/statfacts/html/breast.html (2024).
4. Mahmood, T., Rehman, A., Saba, T., Nadeem, L. & Bahaj, S. A. O. Recent advancements and future prospects in active deep learning for medical image segmentation and classification. *IEEE Access* (2023).
5. Meena, G., Mohbey, K. K., Acharya, M. & Lokesh, K. Original research article an improved convolutional neural network-based model for detecting brain tumors from augmented mri images. *J. Auton. Intell.***6** (2023).
6. Meena, G., Mohbey, K. K. & Kumar, S. Monkeypox recognition and prediction from visuals using deep transfer learning-based neural networks. *Multim. Tools Appl.* 1–25 (2024).
7. Meena, G., Indian, A., Mohbey, K. K. & Jangid, K. Point of interest recommendation system using sentiment analysis. *J. Inf. Sci. Theory Pract.* **12**, 64–78 (2024).
8. Mahmood, T., Rehman, A., Saba, T., Wang, Y. & Alamri, F. S. Alzheimer's disease unveiled: Cutting-edge multi-modal neuroimaging and computational methods for enhanced diagnosis. *Biomed. Signal Process. Control* **97**, 106721 (2024).
9. Loh, H. W. et al. Application of deep learning models for automated identification of Parkinson's disease: a review (2011–2021). *Sensors* **21**, 7034 (2021).
10. Mahmood, T. et al. A brief survey on breast cancer diagnostic with deep learning schemes using multi-image modalities. *IEEE Access* **8**, 165779–165809 (2020).
11. Torabi, M., Rasouli, A. H., Wu, Q. J., Cao, W. & Pourpanah, F. Self-supervised adversarial adaptation network for breast cancer detection. *Eng. Appl. Artif. Intell.* **133**, 108489 (2024).
12. Qian, N. *et al.* Lesion attention guided neural network for contrast-enhanced mammography-based biomarker status prediction in breast cancer. *Comput. Methods Progr. Biomed.* 108194 (2024).
13. Mahmood, T., Saba, T., Rehman, A. & Alamri, F. S. Harnessing the power of radiomics and deep learning for improved breast cancer diagnosis with multiparametric breast mammography. *Expert Syst. Appl.* **249**, 123747 (2024).
14. Mahmood, T. et al. Breast lesions classifications of mammographic images using a deep convolutional neural network-based approach. *PLoS One* **17**, e0263126 (2022).
15. Prusty, S., Patnaik, S., Dash, S. K. & Prusty, S. G. P. Semel-lr: An improvised modeling approach using a meta-learning algorithm to classify breast cancer. *Eng. Appl. Artif. Intell.* **129**, 107630 (2024).
16. Nissar, I., Alam, S., Masood, S. & Kashif, M. Mob-cbam: A dual-channel attention-based deep learning generalizable model for breast cancer molecular subtypes prediction using mammograms. *Comput. Methods Progr. Biomed.* **248**, 108121 (2024).
17. Uddin, K. M. M., Biswas, N., Rikta, S. T. & Dey, S. K. Machine learning-based diagnosis of breast cancer utilizing feature optimization technique. *Comput. Methods Progr. Biomed. Update* **3**, 100098 (2023).
18. Manzo, G. et al. Breast cancer survival analysis agents for clinical decision support. *Comput. Methods Progr. Biomed.* **231**, 107373 (2023).
19. Yuan, H. & Xu, H. Deep multi-modal fusion network with gated unit for breast cancer survival prediction. *Comput. Methods Biomech. Biomed. Eng.* 1–14 (2023).
20. Győrffy, B. Survival analysis across the entire transcriptome identifies biomarkers with the highest prognostic power in breast cancer. *Comput. Struct. Biotechnol. J.* **19**, 4101–4109 (2021).
21. Mahmood, T., Li, J., Pei, Y. & Akhtar, F. An automated in-depth feature learning algorithm for breast abnormality prognosis and robust characterization from mammography images using deep transfer learning. *Biology* **10**, 859 (2021).
22. Varghese, J. Artificial intelligence in medicine: chances and challenges for wide clinical adoption. *Visc. Med.* **36**, 443–449 (2020).
23. Arrieta, A. B. et al. Explainable artificial intelligence (Xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Inf. Fus.* **58**, 82–115 (2020).

24. Wani, N. A., Kumar, R., Bedi, J., Rida, I. *et al.* Explainable ai-driven iomt fusion: Unravelling techniques, opportunities, and challenges with explainable ai in healthcare. *Inf. Fus.* 102472 (2024).
25. Singh, L. K., Khanna, M. & Singh, R. An enhanced soft-computing based strategy for efficient feature selection for timely breast cancer prediction: Wisconsin diagnostic breast cancer dataset case. *Multim. Tools Appl.* 1–66 (2024).
26. Manikandan, P., Durga, U. & Ponnuraja, C. An integrative machine learning framework for classifying seer breast cancer. *Sci. Rep.* **13**, 5362 (2023).
27. UCI. Breast cancer data. https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic (2024).
28. Lundberg, S. A unified approach to interpreting model predictions. arXiv preprint http://arxiv.org/abs/1705.07874 (2017).
29. Setiono, R. Extracting rules from pruned neural networks for breast cancer diagnosis. *Artif. Intell. Med.* **8**, 37–51 (1996).
30. Koyuncu, H. & Ceylan, R. Artificial neural network based on rotation forest for biomedical pattern classification. In *2013 36th International Conference on Telecommunications and Signal Processing (TSP)*, 581–585 (IEEE, 2013).
31. Aalaei, S., Shahraki, H., Rowhanimanesh, A. & Eslami, S. Feature selection using genetic algorithm for breast cancer diagnosis: experiment on three different datasets. *Iran. J. Basic Med. Sci.* **19**, 476 (2016).
32. Naik, A. K., Kuppili, V. & Edla, D. R. Efficient feature selection using one-pass generalized classifier neural network and binary bat algorithm with a novel fitness function. *Soft Comput.* **24**, 4575–4587 (2020).
33. Kıyan, T. & Yıldırım, T. Breast cancer diagnosis using statistical neural networks. *IU-J. Electr. Electron. Eng.* **4**, 1149–1153 (2004).
34. Karabatak, M. & Ince, M. C. An expert system for detection of breast cancer based on association rules and neural network. *Expert Syst. Appl.* **36**, 3465–3469 (2009).
35. Sidey-Gibbons, J. A. & Sidey-Gibbons, C. J. Machine learning in medicine: a practical introduction. *BMC Med. Res. Methodol.* **19**, 1–18 (2019).
36. Obaid, O. I. et al. Evaluating the performance of machine learning techniques in the classification of Wisconsin breast cancer. *Int. J. Eng. Technol.* **7**, 160–166 (2018).
37. Hassan, M. M. *et al.* A comparative assessment of machine learning algorithms with the least absolute shrinkage and selection operator for breast cancer detection and prediction. *Decis. Anal. J.* 100245 (2023).
38. Wani, N. A., Kumar, R. & Bedi, J. Harnessing fusion modeling for enhanced breast cancer classification through interpretable artificial intelligence and in-depth explanations. *Eng. Appl. Artif. Intell.* **136**, 108939 (2024).
39. Chaurasia, V. & Pal, S. Applications of machine learning techniques to predict diagnostic breast cancer. *SN Comput. Sci.* **1**, 270 (2020).
40. Liu, N., Qi, E.-S., Xu, M., Gao, B. & Liu, G.-Q. A novel intelligent classification model for breast cancer diagnosis. *Inf. Process. Manag.* **56**, 609–623 (2019).
41. Abdar, M. et al. A new nested ensemble technique for automated diagnosis of breast cancer. *Pattern Recogn. Lett.* **132**, 123–131 (2020).
42. Sahu, B., Mohanty, S. & Rout, S. A hybrid approach for breast cancer classification and diagnosis. *EAI Endorsed Trans. Scalable Inf. Syst.***6** (2019).
43. Nilashi, M., Ibrahim, O., Ahmadi, H. & Shahmoradi, L. A knowledge-based system for breast cancer classification using fuzzy logic method. *Telemat. Inform.* **34**, 133–144 (2017).
44. Sakri, S. B., Rashid, N. B. A. & Zain, Z. M. Particle swarm optimization feature selection for breast cancer recurrence prediction. *IEEE Access* **6**, 29637–29647 (2018).
45. Ramadevi, G. N., Rani, K. U. & Lavanya, D. Importance of feature extraction for classification of breast cancer datasets—a study. *Int. J. Sci. Innov. Math. Res.* **3**, 763–768 (2015).
46. Murugesan, S., Bhuvaneswaran, R., Khanna Nehemiah, H., Keerthana Sankari, S. & Nancy Jane, Y. Feature selection and classification of clinical datasets using bioinspired algorithms and super learner. *Comput. Math. Methods Med.* **2021**, 6662420 (2021).
47. Idris, N. F. & Ismail, M. A. Breast cancer disease classification using fuzzy-id3 algorithm with fuzzydbd method: automatic fuzzy database definition. *PeerJ Comput. Sci.* **7**, e427 (2021).
48. Rajaguru, H. Analysis of decision tree and k-nearest neighbor algorithm in the classification of breast cancer. *Asian Pac. J. Cancer Prev. APJCP* **20**, 3777 (2019).

Author contributions

S.S. and N.A.W. did conceptualization, methodology, and experimentation and wrote the original draft. S.C. did data curation and analysed the results. N.K. did formal analysis, investigation, validation and supervision. A.M.A. did the review & editing of the manuscript, supervision, and arranged resources.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.