scientific reports

Check for updates

OPEN

A hybrid explainable federated-based vision transformer framework for breast cancer prediction via risk factors

Aymen M. Al-Hejri[1,2,7]✉, Archana Harsing Sable[1,7], Riyadh M. Al-Tam[1,2], Mugahed A. Al-antari[3], Sultan S. Alshamrani[4], Kaled M. Alshmrany[5] & Wedad Alatebi[6]

Breast cancer remains a leading cause of mortality in women, underscoring the need for timely and accurate diagnosis. This paper addresses this challenge by introducing a comprehensive explainable federated learning framework for breast cancer prediction. We evaluate three deep learning approaches in both centralized and federated scenario settings: (1) individual artificial intelligence (AI) models, (2) high-level feature space ensemble models, and (3) a hybrid model combining global Vision Transformer (ViT) and local convolutional neural network (CNN) features. These models are assessed on binary, multi-class, and Breast Imaging Reporting and Data System (BI-RADS) classification tasks using a unique dataset encompassing real-world risk factors. In the federated scenario, we employ three clients with the same approaches as the centralized setting, aggregating their predictions using an AI global model. Explainable AI (XAI) technique is incorporated to enhance AI models' transparency. Our federated learning approach demonstrates superior performance, achieving accuracies of 98.65%, 97.30%, and 95.59% for binary, multi-class, and BI-RADS tasks, respectively. The proposed model, evaluated with a 95% Confidence Interval (CI) and Areas Under Curve (AUC) curves, registers top classifiers with an AUC of 0.970 [0.917–1]. Local Interpretable Model-Agnostic Explanations (LIME) XAI-based federated learning framework offers a promising solution for privacy-preserving and accurate breast cancer prediction in both research and clinical practice.

Keywords Breast cancer, Risk factors, Federated and centralized learning, Ensemble learning, Vision transformer, Explainable artificial intelligence

Breast cancer is the most commonly diagnosed cancer worldwide, accounting for 2.3 million new cases annually, which represents 12.5% of all new cancer diagnoses[1]. It remains one of the leading causes of cancer-related mortality, particularly among women, with an estimated 297,790 new cases of invasive breast cancer projected in the United States[2]. Early detection and classification are essential for improving survival rates, as 90% of cases identified in the early stages are treatable[2]. Early detection relies on recognizing specific symptoms, such as palpable lumps, changes in breast size or shape, and nipple discharge, along with identifying risk factors[3].

Breast cancer risk factors are diverse and include controllable elements like obesity and alcohol consumption, as well as non-controllable factors such as genetic predisposition, early menstruation, and late menopause[4]. Additional factors include age, family history, reproductive history, and postmenopausal hormone therapy[4]. Genetic mutations, such as BRCA1, BRCA2, and PALB-2, significantly increase susceptibility[5]. Comprehensive risk profiling and symptom identification play a vital role in clinical diagnosis and management.

Therefore, numerous techniques based on AI have emerged as transformative tools in breast cancer research, offering the ability to analyze vast datasets and extract meaningful diagnostic insights[6,7]. AI techniques, including

[1]School of Computational Sciences, Swami Ramanand Teerth Marathwada University, Nanded, Maharashtra 431606, India. [2]Faculty of Administrative and Computer Sciences, University of Albaydha, Albaydha, Yemen. [3]Department of Artificial Intelligence and Data Science, Daeyang AI Center, College of AI Convergence, Sejong University, Seoul 05006, Republic of Korea. [4]Department of Information Technology, College of Computers and Information Technology, Taif University, PO Box 11099, 21944 Taif, Saudi Arabia. [5]Institute of Public Administration, P.O.Box 5014, 21944 Jeddah, Saudi Arabia. [6]Department of Statistics, College of Science, Tabuk University, PO Box 741, Tabuk, Saudi Arabia. [7]Aymen M. Al-Hejri and Archana Harsing Sable contributed equally to this work. ✉email: aymen.alhejri@baydaauniv.net

machine learning (ML), deep learning (DL), transfer learning, ensemble learning, and ViT, have demonstrated significant promise in breast cancer prediction[6,8,13–15]. These methods excel in handling diverse datasets, such as imaging data (e.g., mammograms, ultrasounds) and structured data (e.g., WDBC, patient medical histories and risk factors)[9,10]. Among these, federated learning has gained attention for its ability to address challenges in data sharing and privacy[11]. This collaborative learning paradigm trains models across distributed datasets without centralizing sensitive data, ensuring privacy and security[12]. Federated learning effectively handles data imbalances and heterogeneity across nodes while enabling the construction of robust global models. By combining data from multiple sources without breaching privacy regulations, federated learning has proven particularly suitable for healthcare applications[13].

Furthermore, XAI has appeared as a critical tool in making AI systems interpretable and actionable, particularly in healthcare[14]. By enhancing transparency, XAI builds confidence among clinicians and patients, enabling better utilization of AI-generated insights for tasks such as diagnosis and personalized care based on electronic healthcare records[15]. LIME, a prominent XAI technique, has shown potential in real-world applications by offering meaningful interpretations of AI predictions, although its use beyond machine learning research has been limited[16].

In this study, we propose a comprehensive explainable ensemble transformer federated learning framework (CEET-Fed) to enhance breast cancer prediction. This framework integrates federated learning with XAI techniques, offering accurate predictions while maintaining interpretability. Using a unique dataset that combines patient medical histories, risk factors, symptoms, and imaging data, we implement and evaluate three distinct approaches: individual models, ensemble models, and a hybrid ViT-CNN model in centralized learning models, then applied the hybrid ViT-CNN (CEET-Fed) in federated scenario settings. The study includes a detailed comparison across binary, multi-class, and BI-RADS classification tasks. To ensure a robust evaluation, the study conducted experiments in centralized and federated learning scenarios. In the centralized setting, recent machine learning models, including LightGBM, Gradient Boosting, XGBoost, CatBoost, and Random Forest (RF), were utilized[17–21]. These models were ensembled to optimize prediction accuracy using soft voting and bagging techniques. Simultaneously, transfer learning and transformer-based approaches were utilized to extract features from risk factors, which were then combined to enhance overall performance. In the federated learning scenario, models were trained locally on distributed datasets before aggregating the results to update a global model. The results of these experiments were compared to evaluate the relative advantages of centralized versus federated learning approaches. The primary contributions of this study are as follows:

- The paper introduces a novel CEET-Fed framework that integrates federated learning with explainable AI for breast cancer prediction.
- Evaluation of centralized and federated learning scenarios using advanced machine learning and deep learning approaches in three different tasks: binary, multi-class, and BI-RADS classifications.
- The study utilizes a comprehensive dataset that includes real-world risk factors for breast cancer, contributing to the development of practical and accurate breast cancer prediction models.

The remainder of this paper is structured as follows: Sect "Related work" reviews related literature on breast cancer risk factors and AI-based models. Sect "Materials and method" describes the proposed models and methodologies. Sect "Experimental results and discussion" presents experimental results and analysis discussion. Sect "Limitations and future works" discusses the study's limitations and future directions. Finally, Sect "Conclusion" concludes the paper.

Related work

Prediction of breast cancer: risk and clinical data

Several significant risk factors for breast cancer are indicated in its occurrence, including non-modifiable factors such as being female, having a family history of the disease, genetic alterations, menstrual period and menopausal trends, and previous radiation therapy. Additional factors include modifiable factors such as Hormonal replacement therapy, Physical activity, Overweight, Alcohol consumption, or smoking[22,23]. Currently, most breast cancer patients, almost 80%, are individuals aged over 50. Increased Body Mass Index (BMI) is an indicator of the risk factors and presence of breast cancer. The connection between higher BMI and the chance of developing cancer is elucidated by several factors, such as the heightened conversion of androgens to estrogens, the activation of insulin and insulin-like growth factor (IGF) signaling, the abnormal functioning of adipokines, and the presence of chronic inflammation[24]. Regarding the risk and clinical data, Yala, Adam, et al.[18] developed logistic regression (RF-LR) and CNN models to predict breast cancer using risk factors and mammography. RF-LR trained using risk factors, questionnaires from patients, and electronic medical records; however, CNN deep learning for image-only evaluation and a hybrid deep learning model included both images and risk factors. The experiments were conducted on a dataset of 39,571 women. The Hybrid DL, image-only DL, and RF-LR models demonstrated AUC values of 0.70 (95% CI: 0.66, 0.75), 0.68 (95% CI: 0.64, 0.73), and 0.67 (95% CI: 0.62, 0.72), respectively. In addition, Akselrod-Ballin, Ayelet, et al.[25] also suggested a model that combines ML and DL to analyze a connected set of mammography images and electronic health data for a total of 13,234 women. The model achieved an area under the AUC of 0.91 (95% CI: 0.89, 0.93), with a specificity of 77.3% (95% CI: 69.2%, 85.4%) and a sensitivity of 87%. In contrast, the model achieved an AUC of 0.78 based solely on the data. Another study developed a logistic regression model based on risk factors, termed the (RF-LR) mode, aimed at predicting the risk of breast cancer[26]. This model utilized a dataset comprising 88,994 mammograms collected from 39,571 women. The RF-LR model attained an AUC of 0.67 (95% CI: 0.62, 0.72), incorporating deep learning methodologies alongside mammographic and clinical data to improve the accuracy of breast cancer risk predictions.

The authors[27] collected a dataset from the University of Coimbra, which consists of 116 samples. Out of the total number of cases, 64 have been verified to have breast cancer, while the remaining cases are classified as healthy. Age, BMI, Glucose, and Insulin were among the nine clinical factors examined to determine the probability of breast cancer. Three machine learning models, namely LR, RF, and SVM, were utilized. The SVM model showed the highest performance, with an AUC of 95% with a CI of 0.87 to 0.91. The researchers in[28] utilized a CNN-based approach to integrating mammographic examination and clinical data to predict breast cancer risks. Data on risk factors combined with mammography for 23,467 women was collected from electronic health records (EHR). The assessment results were obtained using logistic regression for clinical factors exclusively and a hybrid CNN model combining clinical factors and mammography image data. The AUC for the logistic regression model was 0.624, while the AUC for the hybrid model was 0.654. Conversely, other papers utilized features derived from a digitized image, specifically the Wisconsin Diagnostic Breast Cancer (WDBC) breast cancer dataset[29]. Hamedani-Kar, et al.[6] proposed the utilization of uncertainty quantification (UQ) and DenseNet121 transfer learning to enhance the reliability of uncertainty quantification. In the study[30], the researchers introduced an artificial neural network (ANN) model that utilized WBCD and WDBC datasets, along with feature optimization. Their findings demonstrated an accuracy rate of 99.85% for WBCD and 99.47% for WDBC. Another study took a look at what could increase the chances of getting cervical cancer by using data on risk factors from the Hospital University de Caracas[31]. They pulled out some key features, and with that, their machine-learning models hit an impressive AUC of 0.95. To make sense of the results, they used Shapley additive explanations (SHAP) as a way to explain their findings.

Prediction of breast cancer: transfer learning, ensemble learning, and transformer learning

Transfer learning is a technique that involves leveraging the pre-trained weights of a model to avoid the need for extensive computational resources and effort[32]. Pre-trained models are models that were previously trained to address a similar problem[33]. For example, Addo, Daniel, et al. [34] suggested a model consisting of depth-wise separable convolution, second-order pooling, and multi-head self-attention modules for classification histopathology images. Furthermore, the authors introduced an ensemble learning model[8] that combines three transfer learning models (AlexNet, ResNet, and MobileNetV2) using a mini-DDSM dataset and an ultrasound dataset (BUSI). The results achieved a detection accuracy of 97.75% for identifying malignancy in the mini-DDSM dataset and 94.62% for detecting both abnormality and malignancy. The authors[35,36] employed the AlexNet and Dense-121 models to extract the deepest features to evaluate breast cancer and diabetic retinopathy. Their trials yielded reliable outcomes by utilizing CNN features that were extracted. However, the ensemble approach is a method that combines multiple individual transfer models to improve accuracy, but it incurs higher computational costs compared to using single models[37]. Moreover, the study[38] employed various single models, including GoogleNet, InceptionResNetV2, VGG16, ResNet50, and Xception. They fine-tuned and froze certain layers as non-trainable layers. Subsequently, they merged the high-level features to serve as the backend of the convolutional network with the self-attention ViT. The model obtained a remarkable accuracy of 97.87% on INbreast mammograms. The study[39] presented two ensemble models that utilize five transfer learning models for detecting pneumonia diseases in X-ray images. The models attained a classification accuracy of 98.19% for multi-class classifications. The ViT-based model was applied to the DDSM dataset, and a classification accuracy of 95.80% for multi-class classification was obtained[40]. In addition, the WDBC dataset was used to build a two-layer nested-based ensemble model, which was then applied to the dataset and attained an accuracy of 98.07%[41]. On the other side, a method was created based on the self-supervised version of ViT and knowledge distillation on the BUSI dataset[42]. Through the utilization of knowledge distillation, it is possible to train a single vision transformer to match the performance of an ensemble consisting of a certain set of vision transformers[43]. A stacked ensemble of three ResNet models (ResNet50V2, ResNet101V2, and ResNet152V2) was proposed for the classification of breast cancer masses as malignant or benign, achieving an accuracy between 85 and 99%, based on BI-RADS score (2–6)[44]. Furthermore, the study[45] proposed an ACSNet multi-task learning model that combines an attention mechanism and a multi-scale feature fusion method for breast cancer classification. The model optimizes the information flow between the encoder and decoder, resulting in outstanding performance.

Prediction of breast cancer: federated learning

Federated learning is one of the cutting-edge technologies that has made it feasible to engage in collaborative machine learning on multiple devices[46]. A federated learning framework is a highly promising approach for implementing machine learning on edge devices[47]. It utilizes decentralized ML techniques to create a shared model by leveraging local data and communicates model updates to centralized parameter servers, rather than transmitting raw data[48]. It has demonstrated a great deal of potential and has made tremendous progress[49]. In addition, federated learning possesses significant potential for advanced healthcare applications, as demonstrated in this study[50], which implemented the federated modeling process alongside an analysis of explainability. One example is FRESH, a healthcare system that relies on FL and the sharing of physiological data[51]. Edge computing devices utilize local data to train machine learning models and enhance the accuracy of disease prediction, hence optimizing data processing. For example, Almufareh, M.F. et al.[52] employed collaborative federated learning for detecting breast cancer using deep neural networks while ensuring privacy data. Their model is implemented on the BCW dataset, yielding an accuracy of 97.54% and an F1-score of 97%. Furthermore, the authors in the study[31] offered a deep learning approach for classifying lung cancer at several orders using an ensemble federated learning framework with enhanced accuracy evaluation. The researchers[53] suggested employing the chaotic tuna swarm optimization (CTSO) method and the BreakHis database to implement federated learning. The model achieved an accuracy of 95.68%. Recently, the SHEFL[54] has been proposed as a means of encrypting the decentralized image cancer to achieve state-of-the-art performance while simultaneously guaranteeing privacy through federated learning.

Material and method

The main objective of our approach is to construct and provide a detailed overview of the cutting-edge techniques applied in breast cancer detection models utilizing risk factors and patients' health records within centralized and federated learning scenarios. Initially, the collection of data related to the risk of breast cancer involved the recruitment of women who had just received a diagnosis of breast cancer. These women were selected from the National Cancer Control Foundation (NCCF) in Sanaa, Republic of Yemen. The health records and medical history of each patient were documented to identify and analyze the risk factors labels associated with breast cancer. The diagnosis was established based on a positive mammography or ultrasound medical report, with some cases requiring further review and confirmation through histopathology. Figure 1 provides a detailed workflow outlining the development of the proposed approach, beginning with the acquisition of health record data. Subsequently, incorporate cancer classification based on the medical reports provided by radiologists. Next, the dataset undergoes preprocessing, followed by feature selection and over-sampling techniques to address data imbalance. Ultimately, two distinct scenarios are then employed for training models: centralized learning and federated learning. The utilization of centralized models involves the sharing of data through a central server, which is afterward accessed by other participants. To acquire sufficient data from various clients and diverse risk data, we applied again the proposed hybrid model CEET in a federated learning model with the name CEET-Fed model. FL ensures data privacy and integrity at the individual client level, as opposed to sharing information on a centralized server.

Dataset description

This study involved collecting risk factors and health records in unstructured electronic health records (EHR) format. The institutional review board (IRB) approved the dataset for research purposes under approval number SUIRB-HR-2023-013 and management number SU-2023-013-01 from the National Cancer Control Foundation (NCCF) in Sanaa, Republic of Yemen. Medical experts gathered the dataset from April 2022 to March 2023 with 802 patients: 227 normal, 409 with benign conditions, 98 malignant, and 68 alongside medical radiology reports. In this study, we exclude the cases (i.e., 68 cases) having a lack of medical reports. Figure 2 depicts the dataset distribution per class. Our dataset includes 41 risk factor features classified according to the BI-RADS system[55], which encompasses: (1) Normal, (2) Benign findings, (3) Probably benign findings, (4) Suspicious abnormality, (5) Highly suggestive of malignancy, and (6) Proven malignancy. In diagnosing breast cancer, clinicians focus on key aspects: patient history, symptoms (age, marital status, self-examination, clinical exam results, annual mammograms), family medical history (parental consanguinity, previous cancer occurrences), and prior breast procedures (mastectomy, lumpectomy). Physical examinations assess signs like skin changes, shape or size changes, nipple changes, mass presence, pain, axillary mass, breast itch, or nipple discharge. Other factors are also considered, including BMI, menstrual regularity, contraceptive use, and hormonal replacement therapy.

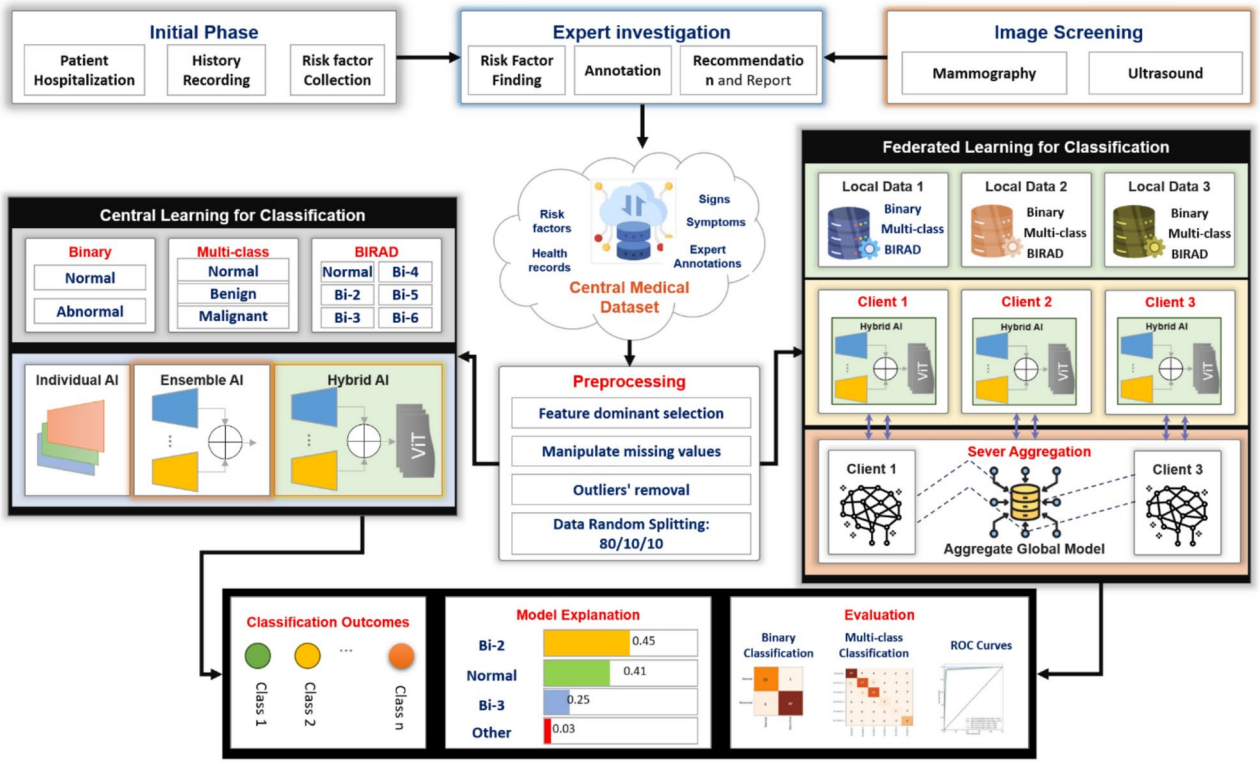

Fig. 1. The abstract of the proposed CEET-Fed framework for predicting breast cancer risk factors is based on a combination of centralized and federated learning scenarios.

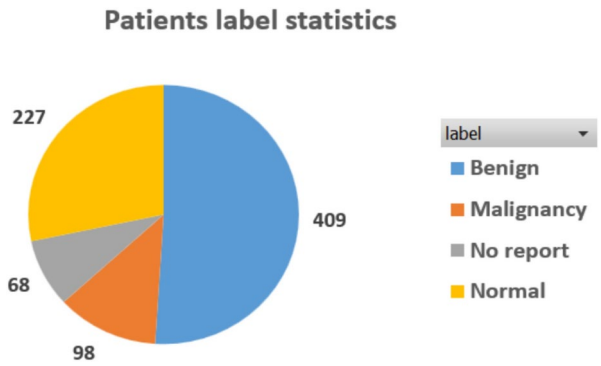

Fig. 2. The number of patients that were collected per label.

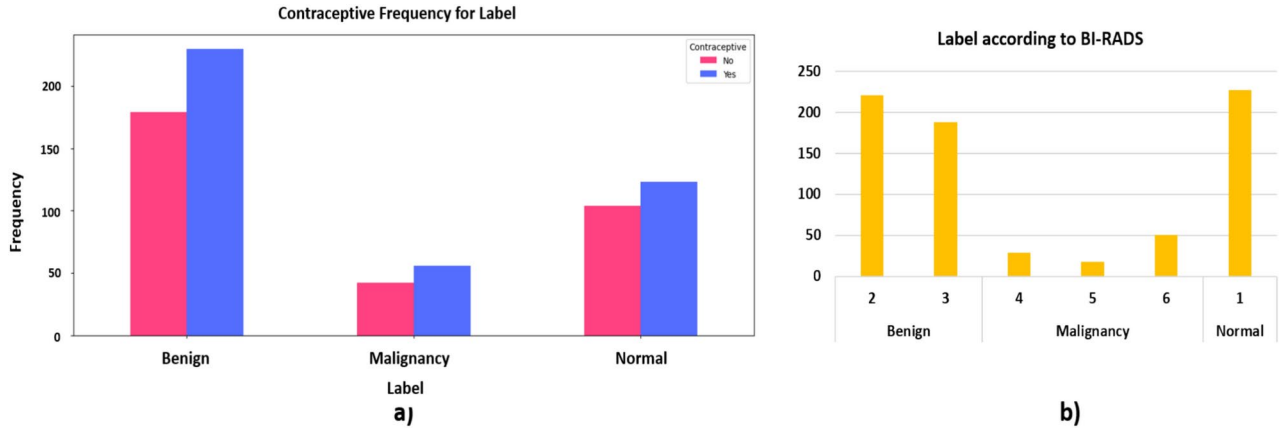

Fig. 3. Data frequency with label: (**a**) The contraceptive frequency for label, (**b**) The BI-RADS frequency for label.

The goal is to compile comprehensive health records, especially medical radiology reports, to identify breast cancer risk factors. These factors help classify breast cancer into two categories (Normal and Abnormal), three categories (Normal, Benign, and Malignant), as well as BI-RADS categories (1–6), as shown in Fig. 3. Moreover, a detailed summary of the breast cancer risk data is provided, as tabulated in Table 1.

Preprocessing

In the first phase of data preprocessing for this study, we conducted a thorough examination of all characteristics to identify any instances of null values and empty rows. Subsequently, we proceeded to address the null values by replacing them with the median value. Furthermore, check for outlier values entered incorrectly by users. For example, some outliers for age or weight were replaced by considering the characteristics of similar patients. Then, the categorical features were encoded to equivalent numeric values using Python library of "OrdinalEncoder"[56]. Finally, the data was normalized and scaled using the function of "MinMaxScaler". This is to scale and mimic all data features to be in a similar range, thereby preventing any potential bias, as discussed in our previous work[57].

Features selection

In the process of feature engineering, it is common to examine characteristics that show zero variation or standard deviation[58]. Subsequently, features showing a weak correlation are eliminated from the investigation, and a random forest technique is employed to determine the most significant features. Table 2 and Fig. 4 present a set of selected features based on their respective levels of significance. Features with values below a threshold of 0.011 are eliminated. This threshold is determined based on the random forest approach, which is known for its low tendency to overfit and its ability to provide a high level of interpretability, ensuring robust and reliable feature selection[59].

2D data representation for AI-based CNN modeling

The tabular data, in which information is structured in rows and columns, is imported into a DataFrame by the Panda library when the data has been organized as a CSV file. Each row in the dataset represents a single patient, with each column corresponding to a specific feature or dimension. The goal is to convert and modify the tabular data to make it appropriate for modeling with CNNs. To prepare the data for training and testing, which have dimensions (X, 32, 32), the tabular data is transformed into two-dimensional forms, as outlined in Algorithm 1.

#	Feature	Values	#	Feature	Values
1.	Age	Numeric	2.	Regular exercises	Boolean
3.	Marital state	Single (0), Married (1), widow (2), and divorced (4)	4.	Smoking history	Boolean
5.	Self-examination	Boolean	6.	Chewing Qat	Boolean
7.	Clinical examination	Boolean	8.	Shama uses	Boolean
9.	Annual mammography	Boolean	10.	Shisha uses	Boolean
11.	Indication	Check-up, Diagnostic, Follow up	12.	Weight	Float
13.	Skin change	Boolean	14.	Length	Float
15.	Shape or size change	Boolean	16.	BMI	Float
17.	Nipple change	Boolean	18.	Regular menstruation	Boolean
19.	Presence of mass	Boolean	20.	Absent menstruation	menopause, hysterectomy, none
21.	Presence of pain	Boolean	22.	Contraceptive	Boolean
23.	Axillary mass	Boolean	24.	Hormonal replacement	Boolean
25.	Presence of itch	Boolean	26.	Other hormonal treatment	Tamoxifen, Steroids, Other, no
27.	Nipple discharge	Boolean	28.	BI-RADS	(1) Normal (2) Benign finding (3) probably benign findings (4) Suspicious abnormality (5) Highly suggestive of malignancy (6) proven malignancy
29.	Family history	Family member, family member with ovarian cancer, none	30.	Regular exercises	Boolean
31.	Parents' consanguinity	First degree, second degree, none	32.	Smoking history	Boolean
33.	Other cancer	Uterus, colon, lymphoma, ovary, Thyroid, none	34.	Chewing Qat	Boolean
35.	Previous biopsy	FNAC, biopsy, open Biopsy, none	36.	Radiotherapy	Boolean
37.	Previous breast operation	Mastectomy, Lumpectomy, Abscess drainage, Breast reconstruction, others, none	38.	Number of pregnancies	Numeric
39.	Menarche age	Numeric	40.	Label	Normal, Benign, Malignant
41.	Late menopause	Boolean			

Table 1. Comprehensive list of breast cancer risk factors and corresponding dataset features.

#	Clinical features	Significance measure	#	Clinical features	Significance measure
1.	Age	0.108830	13.	Clinical examination	0.021706
2.	BMI	0.105780	14.	Family history	0.020616
3.	Weight	0.097979	15.	Presence of itch	0.019027
4.	Length	0.085578	16.	Indication	0.018681
5.	Number of pregnancies	0.065530	17.	Contraceptive	0.018022
6.	Presence of mass	0.065469	18.	Shisha uses	0.017340
7.	Menarche age	0.064073	19.	Presence of pain Hormonal replacement	0.016551
8.	Chewing Qat	0.045273	20.	Shape or size change	0.013659
9.	Smoking history	0.028118	21.	Nipple discharge	0.012655
10.	Marital state	0.026338	22.	Axillary mass	0.011826
11.	Parents' consanguinity	0.022270	23.	Absent menstruation	0.011447
12.	Regular menstruation	0.021769			

Table 2. The selected features using the RF approach, which ranked by their relative significance.

Client selection

In the federated learning framework, clients represent distinct data sources, such as organizations, institutions, or hospitals, that collaborate while maintaining data privacy. The selection of clients is a crucial factor that influences model performance and generalizability[60].

In real-world applications, the number of clients can vary depending on multiple factors, including data availability, computational resources, and communication infrastructure. For instance, prior studies, such as[61], have experimented with different numbers of clients (e.g., 5, 10, and 15) on a large IoT dataset, where the 10-client setup demonstrated slightly better performance compared to others. In our study, we selected three clients based on the following considerations:

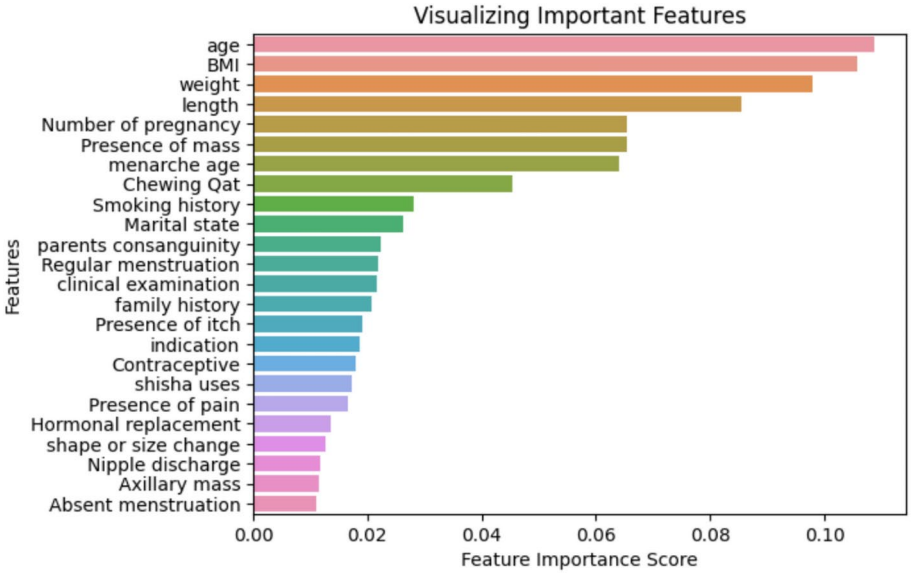

Fig. 4. The most feature significance values that obtained by the utilization of a random forest technique.

Start:

Declare function reshape_2d ($dataX, n = 32, m = 32$):

Declare list Xt

for each row r in $dataX$:

Initialize $z1$ as NumPy zeros with 1024 length

Assign $z1$ from 1st index to length $dataX$ with $dataX$[r]

Append Xt by $z1$

End for

Convert Xt list to NumPy array

Reshape Xt to length of $dataX,n,m$

$return\ Xt$

End

Algorithm 1. Resizing the tabular data from 1 to 2D for CNN operation.

- The dataset used in our study contains diverse breast cancer risk factors, and splitting it among three clients allows for a balanced and representative distribution while maintaining sufficient data volume for effective training.
- The federated learning setup requires continuous model aggregation and synchronization between clients and the central server. Given our available infrastructure, three clients provided an optimal balance between performance and efficiency.
- In practical medical applications, federated learning often involves a limited number of institutions collaborating while ensuring privacy. Our approach mirrors such realistic scenarios where a small but meaningful number of clients participate in federated training.

Data splitting for training, validation, and testing

After the dataset is carefully prepared, we randomly split it per class into 80% for training, 10% for validation, and 10% for testing. The splitting strategy is applied similarly for both centralized and federated learning scenarios. Tables 3 and 4 summarize the dataset distribution after splitting per class over three classification tasks (binary, multi-class, and BI-RADS) for the centralized and federated learning scenarios, respectively. To address the class imbalance, especially in the training sets of all classification tasks, we implement the synthetic minority oversampling technique (SMOTE)[62] to balance the dataset, with a particular focus on the majority class.

#		Binary		Multi-class			BI-RADS					
Class	All	N	Ab	N	B	M	N	Bi-2	Bi-3	Bi-4	Bi-5	Bi-6
Total records	734	227	507	227	409	98	227	221	188	29	18	51
A training set 80%	587	177	410	177	336	74	177	181	155	22	14	38
*SMOTE training set	–	410	410	336	336	336	181	181	181	181	181	181
Validation set 10%	73	24	49	24	37	12	24	21	16	5	3	4
Testing set 10%	74	26	48	26	36	12	26	19	17	2	1	9

Table 3. Data splitting for training, validating, and testing in the centralized scenario for three classification tasks: binary, multi-class, and BI-RADS. N, Ab, B, and M indicate the class labels for normal, abnormal, benign, and malignant, respectively. *Data balancing of the training set using the synthetic minority over-sampling technique (SMOTE).

#	Binary		Multi-classes			BI-RADS					
Class	N	Ab	N	B	M	N	Bi-2	Bi-3	Bi-4	Bi-5	Bi-6
Total records	227	507	227	409	98	227	221	188	29	18	51
Client 1	80	160	78	139	28	72	67	62	12	8	24
Client 2	63	169	74	131	40	83	68	69	6	5	14
Client 3	84	178	75	139	30	86	86	57	11	5	11

Table 4. Data distribution across clients in the federated learning scenario, showing the splitting of training, validation, and testing datasets for each classification task (binary, multi-class, and BI-RADS).

Centralized learning scenario

In the centralized scenario, all the data is gathered and kept in a centralized location, where it is utilized to train a singular global model. For the centralized scenario, the training procedure usually takes place on a centralized server or a high-performance computing machine[63]. In the context of this study, the dataset was employed exclusively inside one group and afterward divided evenly for all training purposes. In this setting, we incorporate conventional and advanced deep-learning classification techniques. For conventional machine learning techniques, individual models are trained and combined using ensemble methods to classify breast cancer across various tasks, including binary, multi-class, and BIRAD classifications. For the deep learning approaches, three AI model frameworks are designed to address all classification tasks: (1) individual AI models, (2) high-level feature-space ensemble or fusion strategy, and (3) the hybridization strategy combining the benefits of both CNN and transformer-based techniques. The next sections present a detailed description of each technique.

Conventional machine learning models: individual and late ensemble fusion

In the centralized learning scenario, we implemented seven state-of-the-art machine learning models, namely LightGBM, gradient-boosting[17], XGBoost[18], Catboost[19], Random Forest[20], Voting Ensemble[21], and BaggingClassifier[64]. These models were initially applied individually to predict breast cancer based on risk factors. Ensemble learning, a technique that combines many base models to get an optimal outcome, was also employed with the help of a soft-voting classifier[65]. This ensemble approach combined the three highest accuracy models, XGBoost, Catboost, and Random Forest.

Deep learning models: individual, ensemble, and hybrid AI models

Transfer learning models are deep-learning models that have gained knowledge and have been pre-trained using a different dataset. They have demonstrated the ability to achieve significantly higher accuracy compared to untrained CNN models. The risk breast cancer dataset for this study is trained with three pre-trained CNNs: VGG16, ResNet50, and Xception. The classification output from these grid models was fed into a feed-forward neural network, which briefly described the model architecture and hyperparameters. Some layers of these models are frozen and tuned, following the approach used in our earlier work[38]. The dimensions of the input data are represented by reshaping the selected features and rows from a CSV file after splitting them based on the size of the training set and testing per approach. The reshaped dimensions are done to be suitable for modeling purposes with CNNs as mentioned in Algorithm 1.

Transfer learning models VGG16, developed by Simonyan et al. in 2014, is a popular model for image classification, identification, and segmentation applications[66]. This model is a deep convolutional neural network architecture of 16 layers. Images are classified using the VGG-16 architecture, starting at 224×224 pixels and subsequently scaled to 227×227 pixels. The truncated VGG-16 model architecture for this work involves fine-tuning from layer 17 to the final layers, which are trainable, while the remaining layers are untrainable. ResNet50, launched in the deep residual network, won first place in the ILSVRC 2015 classification challenge. It is designed with deep learning through residual networks and skip connection blocks. The pipeline magnitude of the

ResNet50 architecture corresponds to the 2D version of ResNet50. The dimensionality of 224 × 224 × 3 images is reduced to 112 × 112 × 3. In contrast to VGG-16, the ResNet50 architecture has pooling layers in its second, fifth, and eighth layers, resulting in a more significant decrease in the image's width and height[67]. The truncated ResNet50 model architecture for this study is fine-tuned from layer 123 to be trainable. Xception is a widely utilized deep learning convolutional neural network design. The Xception architecture is seen as an improved evolution of the Inception model architecture. The architecture is engineered to classify input images across five dimensions, yielding output images of dimension 2048 × 1. Beginning with 299 × 299 × 3 colored images, the spatial dimensions of the images are decreased in each convolutional block[68].

Ensemble learning model For the concatenate ensemble model, instead of using model predictions alone to train a second model, the full model output is used as input features to the second model. The output of models can be the raw probabilities or features extracted intermediate to the final prediction. Thus, every model in the ensemble sees the same data as the second model, but its representation of the data differs based on the architecture and initial random weights[69]. A concatenate ensemble model can capture the strengths of the different models. An obvious pitfall of a concatenation ensemble model is that if it is built on too many complex models, then it may simply learn to choose the best model for predicting any particular item and return model outputs instead of learning to combine the models. To prevent this, deep learning concatenate models typically have a large, fully connected layer below the output layer to force meaningful learning[70]. To enhance the effectiveness of breast cancer risk classification, an ensemble approach that amalgamates predictions from multiple models has been explored in this work. Specifically, the ensemble is constructed by merging two standalone transfer learning models—VGG16 and ResNet50—that are trained from scratch for feature extraction. These models are trained using risk factors and health records, which are recorded along with radiologist medical reports, as shown in Fig. 5.

Transformer learning model In addition, the transformer is a type of deep learning model that uses the self-attention mechanism to assign various levels of priority to specific elements of the input data and is conveyed via an encoder-decoder approach. In the context of breast cancer risk classification, transformers are used to fully leverage the heterogeneous feature representations learned by multiple models, thereby enhancing the overall classification performance. A transformer-based model, which utilizes an attention mechanism to effectively fuse multiple feature representations, is introduced. A transformer encoder architecture is designed to learn a set of latent representations, which are then jointly processed using a multi-headed self-attention mechanism. It is used to obtain a fused representation that captures both the complementary and synergistic information between diverse features[71]. The transformer encoder consists of three main components, which are transformer input, multi-headed self-attention mechanism, and feed-forward neural networks. The ViT encoder produces two outputs from a given input image: the full input token embedding vector across all layers and the attention weights of all heads. The self-attention mechanism can generate a representation of a single input sequence by establishing connections between different positions within the sequence[39,72]. Meanwhile, computed attention functions are stored in the attention matrix, which also includes the set of queries (Q), keys (K), and values (V) as computed in this equation:

$$Attention\,(Q, K, V) = softmax\left(\frac{Qk^2}{\sqrt{d_k}}\right)v \tag{1}$$

Given an input with a dimension d_k for queries and keys and dimension d_v, the dot product of all queries with keys is calculated by normalizing each by $\sqrt{d_k}$ and applying SoftMax to determine the weights for the value pairs.

Multi-headed attention enables the model to concurrently focus on inputs from many representation subspaces across different locations. The calculation for this mechanism is expressed in Eq. 2.

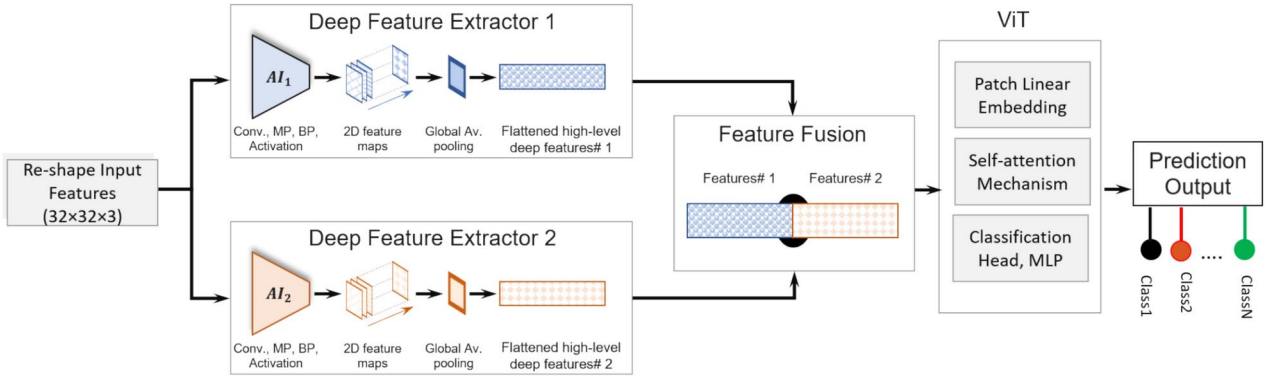

Fig. 5. The proposed hybrid features fusion ensemble-based with transformer ViT model for both centralized and federated learning scenarios.

Variables	Description
Number of Clients	3 number of participants (3 clients)
Data splitting per clients	Splitting data for training (80%), validation (10%), and testing (10%) per client
Number of labels	Classification tasks: binary (N, Ab), multi-class (N, B, M), and BI-RADs (N, Bi-2, Bi-3, Bi-4, Bi-5, Bi-6)
learning_rate	0.001
Batch_size	32
Epochs	50
Rounds	10
Federated averaging process	tensorflow_federated.learning.build_federated_averaging_process(createModel, client_optimizer_fn)

Table 5. The parameters of the CEET-Fed model to update weights between clients and server.

Start:
Initialize the global model parameters θ
Declare list of clients C_i
for each round r do
  for each client k in C_i
    Training local model of client k by CEET-Fed model
    Sending local model of k to Server S
    Retrieve global model parameters θ from Server S
    Update the local model of k
    end for
  Compute global model update $A\theta = Average\ (A\theta i\ for\ each\ k\ in\ Ci)$
  Update global model θ
end for
return global model parameters θ
End

Algorithm 2. Federated learning scenario based on CEET-Fed framework.

$$MultiHead\,(Q, K, V) = Concat\,(head1, \ldots\ldots, headh)\,W^o \tag{2}$$

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right), \tag{3}$$

The encoded structure comprises the self-attention network and the MLP block, along with a normalization layer and residual connections that are directly utilized within each block. An example of a feed-forward neural network model is the multilayer perceptron (MLP), which includes both dense and dropout layers.

Federated learning scenario

In the centralized learning scenario, the proposed hybrid model-based ensemble and ViT achieved the highest accuracy. As a result, we retrained the CEET-Fed model by implementing all three approaches within the FL scenario. Federated learning is well recognized as a decentralized methodology within the field of machine learning, wherein the training of global models is achieved through the collaborative participation of multiple clients. The protection of data privacy is achieved by employing local models that are contributed by participants, instead of exchanging private local data. However, the effectiveness of federated learning is highly dependent on the number of participants and the level of their contributions[73]. The participants proceeded to modify their local models to fulfill the requirements of the assigned rounds. This method offers significant benefits in several dispersed learning contexts. In the context when K clients are participating in federated learning with a shared objective of collectively training a model, the process involves the distribution of the global model to the clients at each iteration. Subsequently, the clients independently train the model using their local data. Once the training at the local level has concluded, the model parameters are transmitted by each client to the central machine. Subsequently, the global model is constructed by combining the model parameters obtained from each client. The proposed CEET-Fed employs again in federated learning scenarios using the TensorFlow 2.3.4 and TensorFlow Federated 0.17.0 libraries to simulate the prediction of breast cancer risk across multiple clients. The model is utilized as a localized model for each of the three clients, with each client having their dataset, as indicated in Table 5. The dataset is partitioned for each client, training each of the models independently for 50 epochs, then updating the weights between the local models and the server in 10 rounds. The pseudo-code

for the federated learning procedure is depicted in Algorithm 2. The parameters of the proposed model in a federated learning scenario are outlined in Table 5.

For the local training on each client, each client k, the client has a local dataset D_k. In each round, the client performs local gradient descent to update the local model $w\text{CEET}_k$. The update is done as follows:

$$wCEET_k^{t+1} = wCEET_k^t - n\nabla L\left(wCEET_k^t, D_k\right) \tag{4}$$

where $wCEET_k^{t+1}$, is the model weights for client k at round t, n is the learning rate,$L\left(wCEET_k^t, D_k\right)$ is the loss function evaluated on the client D_k. and ∇ is the gradient of the loss function[74].

After all the clients complete their local training, the server aggregates their updated models. This is done by computing a weighted average of the model updates from all clients. If there are K clients and each client k has a dataset size of n_k, the updated global model $wCEET_k^t$ at round t is computed as:

$$wCEET^{t+1} = \sum_{k=1}^{K} \frac{n_k}{N} wCEET_k^{t+1} \tag{5}$$

where $wCEET_k^{t+1}$, is the updated model, N is the total number of data points, and $\frac{n_k}{N}$, is the weight assigned to each client. Finally, repeat the training till reaches the predefined number of rounds.

Explanation generation using XAI model

The research field of XAI models is becoming increasingly relevant[75]. With a rising number of AI and ML models, there is a growing demand for them to be understandable to the public and specialists in their respective fields. The LIME approach is a local explanation method that operates in a black-box setting[76]. LIME is an increasingly popular technique to perform the interpretability of any kind of ML model. LIME explains one ML prediction at a time, by learning a simple linear model around the prediction. More precisely, it allows the generation of local explanations of any ML model. The process can be formulated as an optimization problem, which LIME solves to generate an explanation. The output of LIME is an explanation object that includes a paired list of feature contribution values and feature indexes. The list is sorted in descending order based on the feature values. By rearranging this list according to the feature index, the feature contribution values can be plotted against the program location. The feature contribution values will range from -1 to 1, where positive feature contribution values represent the abnormal class, and negative feature contribution values represent the normal class. A LIME explanation applied by the lime library, lime, Lime_tabular, and LimeTabularExplainer parameters with training, classes, and kernel[77].

Evaluation

The overall performance of the proposed models was further evaluated by utilizing the test data, considering the following criteria for detection. Accuracy is a metric used to measure the number of correct classifications out of the total number of entries, as defined by the following formula,

$$Accurac\,(ACC) = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

Precision is the proportion of true positive predictions out of all positive predictions made by the model. This can be computed utilizing,

$$Precision\,(PRE) = \frac{TP}{TP + FP} \tag{7}$$

Recall (also known as sensitivity (SEN) or true positive rate) is the proportion of true positive predictions out of all the actual positive instances in the testing set. It indicates the total number of samples that should have been diagnosed as malignant. Mathematically, it is expressed as,

$$Recall(SEN) = \frac{TP}{TP + FN} \tag{8}$$

The F1-Score is a metric that measures the ratio between Precision and Recall by calculating their Harmonic Mean.

$$F1 - Score = \frac{2Precision \times Sensitivity}{Precision + Sensitivity} \tag{9}$$

AUC is a summary measure of the ROC curve that indicates the classifier's ability to distinguish between classes. If the AUC is high, the model can reliably separate positive from negative samples[78]. The ROC curve can be thought of as a learning aid for the proportions of false positives and true positives. The ROC curve provides insight into evaluating the costs and benefits of the classifier. To quantify how often false positives occur, we can compare the number of misclassified negative samples to the total number of negative samples.

A 95% Confidence Interval [95% CI] is calculated for the AUC of the ROC curve allows one to infer the range of values that the AUC would take from the results obtained from the data. It is the repeated sampling from a population that follows a normal distribution, each time calculating the interval in a certain way[79]. When

x% of the multiple intervals obtained at this time contain the population mean within the interval, the interval obtained by this method is called the x% confidence interval. *X*% confidence interval of the ROC-AUC *C(x)* is,

$$C\left(x\right)=A+\mathrm{z}\left(\frac{1-x}{2}\right)SE\left(A\right), \tag{10}$$

where A is AUC, z(a) is a function that returns the upper a% point of the normal distribution, and it will be 1.96 when determined *x = 0.95,*

Standard Error of AUC (A) can be calculated as

$$SE\left(A\right)=\sqrt{\frac{A\left(1-A\right)+\left(n_p-1\right)\left(Q_1-A^2\right)+\left(n_N-1\right)(Q_2-A^2)}{n_P n_N}}, \tag{11}$$

where:$\mathrm{q1}=\ \mathrm{AUC}/(2-\mathrm{AUC})$, $\mathrm{q2}=2*\mathrm{AUC2}/(1+\mathrm{AUC})$, n_P is the number of cases that should have truly positive results, and n_N is the number of cases that should have truly negative results. Q1, Q2 is a probability value calculated from the test that produced the ROC curve.

Execution environment

The test was carried out using an MSI GS66 laptop which has the following features: an Intel Core i7 11th generation (11800H) processor, 32 GB of RAM, a 2 TB SSD NVMe storage, and an RTX 3080 graphics card with 16 GB of memory. The experiments mentioned in this article were performed using Python 3.10 on Windows 11, with the use of the Keras and TensorFlow backend libraries.

Experimental results and discussion

The results of evaluating the risk factor-based models' prediction are reported for both centralized and federated scenarios with three approaches: binary, three classes, and BI-RADS classification. Various models are employed utilizing ML models, ensemble transfer learning and transformer models, and federated models to predict breast cancer according to breast cancer risks and health records.

Centralized learning scenario

In the centralized scenario, we employ three approaches: (1) advanced AI models (individual) such as LightGBM, GradientBoosting, XGBoost, Catboost, Random Forest, Voting Ensemble, BaggingClassifier; (2) ensemble transfer learning for VGG16, Resnet50, and Xception; and (3) a hybrid transformer model for predicting breast cancer using patient risk factors and health records.

The proposed CEET ensemble and ViT model architecture combine VGG16 and ResNet50 and a transformer encoder. It considers the 1D time-series risk score vector as input, resulting in a unique architecture. The performance of the individual models is presented first, followed by the evaluation of their ensemble and the proposed model. The individual models have their unique strengths when applied to various data types, influencing their prediction scores. As a result, an ensemble of the individual models with ViT is proposed to alleviate the limitations of the individual models and achieve improved prediction performance. Three classification tasks are employed based on radiologist evaluation reports to classify breast cancer into binary, multi-class (normal, benign, and malignant), and BI-RADS classifications.

Binary classification task

The evaluation prediction results for all trained models using the binary approach are shown in Table 6. In the binary approach, the RF model produced the highest classification results among the stand-alone ML models, achieving an overall accuracy of 91.89% and an F1-score of 91.72%. Catboost and LightGBM attained a predicted

Strategy	Models	ACC	SEN	PRE	F1-score	AUC
Individual model	LightGBM	90.54	90.54	90.65	90.58	90.06
	GradientBoosting	87.84	87.84	87.79	87.65	85.34
	XGBoost	89.19	89.19	89.12	89.08	87.26
	Catboost	90.54	90.54	90.49	90.50	89.18
	Random forest	91.89	91.89	92.17	91.72	89.34
Late ensemble learning	Voting Ensemble	91.89	91.89	91.88	91.81	90.22
	BaggingClassifier (RF)	93.24	93.24	93.38	93.14	91.27
Individual model	VGG16	91.89	91.89	92.17	91.72	89.30
	ResNet 50	94.59	94.59	95.01	94.48	92.30
	Xception	90.54	90.54	91.74	90.13	86.50
Ensemble AI	Ensemble learning	94.59	94.59	95.01	94.48	92.30
Hybrid AI	Proposed hybrid CEET-Fed: Ensemble + ViT	97.30	97.33	97.28	97.32	97.00

Table 6. Centralized scenario (binary task): evaluation classification overall results (%) for each strategy using conventional and deep learning approaches.

performance accuracy of 90.54%, which was the second-highest performance. In contrast, GradientBoosting yielded the lowest results, with an accuracy of 87.84%. The voting ensemble learning model combines the RF and Catboost algorithms, resulting in the same level of accuracy as the stand-alone models. However, it demonstrated an improvement in the AUC metric, reaching a value of 90.22%. The baggingClassifier model got superior performance compared to stand-alone models, attaining an accuracy rate of 93.24% and an F1-score of 93.14%. The ensemble transfer learning approach, which combines the high-level features of stand-alone transfer learning models such as VGG16 and ResNet50, yielded even better results, with an accuracy of 94.59% and an F1-score of 94.48%. The proposed hybrid CEET-Fed framework with ensemble and transformer encoder outperforms the stand-alone models. It achieves an accuracy of 97.30% and an AUC of 97.00%.

Multi-class classification task

For the three-class approach, the benign class has the most instances in the training set, with 336 records after splitting. To balance the dataset, SMOTE oversampling was used, resulting in a training set of 1008 rows. The RF model achieved the highest classification performance compared to other machine learning models, with an accuracy of 93.24% and an AUC of 94.24%. The XGboost model achieved a predicted performance accuracy and F1-score of 91.89%. The remaining ML models produced comparable outcomes, but the baggingClassifier model exhibited greater performance compared to the ML models used individually, reaching an accuracy rate of 94.59% and F1-score of 94.56%. The VGG16 and Resnet50 models, when employed for transfer learning, achieved an accuracy of 94.59% and 90.54%, respectively. Afterward, the high-level features were combined without including the classification layers. The ensemble transfer learning model and the hybrid proposed CEET-Fed model achieved outstanding accuracy results, with an accuracy of 95.95% and an F1-score of 95.96% for both models. Table 7 shows the evaluation metrics results of all machine learning, transfer learning, ensemble learning, and transformer encoder models in a centralized scenario with training on risk-factors data.

BI-RADS classification task

The BI-RADS classification system categorizes types of breast cancer into six classes: (1) Normal, (2) Benign finding, (3) Probably benign findings, (4) Suspicious abnormality, (5) Highly suggestive of malignancy, (6) Proven malignancy. The BI-RADS feature provides the label for this classification approach. Among the types of cancer in our dataset, malignant cases are the least common. This category includes a total of 98 cases, which are split into three subcategories: suspected abnormality, suggestive of malignancy, and proven malignancy. Oversampling is used to address the imbalance in the training set, resulting in a total of 1086 rows regarding the majority class using SMOTE. The findings of this approach closely align with the outcomes of prior approaches. The ML models Catboost and RF demonstrated the highest classification performance, achieving accuracy rates of 94.59% and 95.95%, respectively. The baggingClassifier model achieved equivalent accuracy performance in comparison to the RF model with 95.95% accuracy. The transfer of deep learning, ensemble learning, and the proposed hybrid transformer models yielded somewhat reduced accuracy compared to previous approaches due to increasing the number of labels. Both the proposed ensemble learning and hybrid transformer models attained a maximum accuracy of 93.24%, surpassing the performance of the stand-alone transfer models. The Resnet50 model attained an accuracy of 91.89% and an F1-score of 91.84%. VGG16 achieved an accuracy of 90.54% and an F1-score of 90.50%. The ensemble transfer learning model combines both Resnet50 and VGG16 architectures to concatenate their high-level features and then implement them in an ensemble and hybrid CEET-Fed model as shown in Table 8.

The confusion matrix of the most optimal centralized scenario models for every approach is displayed in Fig. 6. In the binary approach, the hybrid transformer model made incorrect predictions in only two instances: one normal case and one abnormal case. However, the 3-class approach accurately predicted two instances as benign and one case as malignancy. The BaggingClassifier model achieved the best accuracy in predicting all labels in the BI-RADS classification, except for BI-RADS 2 and BI-RADS 3, where it had two and one wrong prediction, respectively.

Strategy	Models	ACC	SEN	PRE	F1-Score	AUC
Individual model	LightGBM	90.54	90.54	90.52	90.51	91.61
	GradientBoosting	86.49	86.49	86.52	86.44	88.72
	XGBoost	91.89	91.89	91.89	91.89	92.93
	Catboost	90.54	90.54	90.70	90.51	92.36
	Random Forest	93.24	93.24	93.32	93.26	94.24
Late ensemble learning	Voting Ensemble	90.54	90.54	90.53	90.50	92.12
	BaggingClassifier (RF)	94.59	94.59	94.70	94.56	95.01
Individual model	VGG16	94.59	94.59	94.70	94.56	96.00
	ResNet 50	90.54	90.54	91.84	90.58	93.30
	Xception	85.14	85.14	85.23	85.16	88.30
Ensemble AI	Ensemble learning	95.95	95.95	96.01	95.96	97.30
Hybrid AI	Proposed hybrid CEET-Fed: Ensemble + ViT	95.95	95.95	96.01	95.96	97.30

Table 7. Centralized scenario (multi-class task): Evaluation classification overall results (%) for each strategy using conventional and deep learning approaches.

Strategy	Models	ACC	SEN	PRE	F1-score	AUC
Individual model	LightGBM	90.54	90.54	90.56	90.52	93.73
	GradientBoosting	91.89	91.89	91.88	91.90	94.64
	XGBoost	93.24	93.24	93.80	93.23	95.58
	Catboost	94.59	94.59	94.55	94.61	96.43
	Random Forest	95.95	95.95	96.02	95.95	97.34
Late ensemble learning	Voting Ensemble	94.59	94.59	94.55	94.61	96.43
	BaggingClassifier (RF)	95.95	95.95	96.02	95.95	97.34
Individual model	VGG16	90.54	90.54	91.05	90.52	95.80
	ResNet 50	91.89	91.89	92.05	91.84	96.30
	Xception	87.84	87.84	88.55	87.44	90.70
Ensemble AI	Ensemble learning	93.24	93.24	93.80	93.23	95.58
Hybrid AI	Proposed hybrid CEET-Fed: Ensemble + ViT	93.24	93.24	93.80	93.23	95.58

Table 8. Centralized scenario (BI-RADS task): Evaluation classification overall results (%) for each strategy using conventional and deep learning approaches.

The classification outcomes for all tasks include binary, multi-class, and BI-RADS, based on individual AI models, ensembles, and hybrid approaches documented during validation evaluations. Presenting the training convergence for both scenarios and all approaches would be cumbersome for readers and render the text disorganized. Figure 7 illustrates the learning convergence of the optimal model among individual AI models and the proposed hybrid model. The convergence rates, regarding accuracy and loss function, indicate that the models' learning process occurs without overfitting.

According to the result of AUC with 95% CI as shown in Table 9 for all AI approaches and tasks, the proposed hybrid AI model is achieved the best results for all classification tasks with 0.970 [0.917–1], 0.956 [0.900–1], and 0.970 [0.929–1] for binary, multi-class, and BI-RADS classifications, respectively.

Furthermore, the time complexity of various machine learning and deep learning models must be carefully considered. Table 10 presents a comprehensive analysis of the proposed ensemble and hybrid strategies employed in this study. Individual machine learning models demonstrate a balance between computational efficiency and accuracy, with LightGBM being the fastest (0.17 s, 90.54% accuracy) and RF offering the highest accuracy at a slightly higher time cost of 0.48 s. Late ensemble methods, such as the Voting Ensemble, provide better accuracy with a moderate time complexity of 1.34 s, while BaggingClassifier achieves the same peak accuracy with a significant computational cost of 17.19 s. Ensemble AI and hybrid strategies, such as the proposed hybrid CEET-Fed: Ensemble + ViT, yield competitive accuracy but are the most computationally intensive, with time complexities of 474.24 and 592.24 s, respectively.

Federated learning scenario

In the federated learning scenario, the proposed CEET-Fed model is also trained using all three of the aforementioned approaches. The training incorporated the local models of three clients undergoing independent training without any sharing of data. The training procedure consists of ten iterations of weight updates to the server by three clients, as illustrated by the data splitting in Table 4 per client through the three approaches. Each model is trained for 50 epochs and then updated the weights between the local models and the server in 10 rounds. Table 11 displays the evaluation performance outcomes of the proposed CEET-Fed in a federated learning scenario. The proposed model in the binary approach attained an accuracy rate of 84.59% in the initial round and improved to 98.65% in the tenth round, as shown in Fig. 8, surpassing the performance of the centralized models. The purpose of the FL scenario is to address issues related to privacy and the uneven distribution of breast risk data. Moreover, an increase in the number of repetition rounds leads to a significant improvement in the accuracy of classification models. Nevertheless, in the three-class approach, the model initially achieved an accuracy rate of 73.61% and progressively improved with subsequent rounds of weight updates until it reached 97.30%. The BI-RADS classification did not surpass the accuracy of a centralized model. However, the model achieved an accuracy of 71.82% in the first round and rapidly increased with each subsequent round, reaching 95.95% accuracy. The objective for deploying these AI models on breast cancer risk factors and clinical data is not to propose them as alternatives for digital mammography, but rather as a noninvasive and economical test that can be easily integrated into routine analyses alongside radiological images.

Explaining the predictions of AI classifiers

The results presented in Fig. 9 illustrate the application of the LIME method to explain the predictions of the CEET-Fed model across three distinct breast cancer classification tasks: binary classification, multi-class classification, and BI-RADS task classification. Each subfigure (a, b, and c) showcases how specific features contributed to the model's predictions for individual cases. The prediction probabilities are assigned to classes according to various classification approaches. LIME explains the first assignment of likelihood. To make the forecast, the probability values are compared to the actual class of the target variable. The study included the case with ID 12 for testing interpretation of prediction, this case belonged to the benign class and was considered in the testing set for all approaches. The model provided a higher probability to the abnormal, benign, and

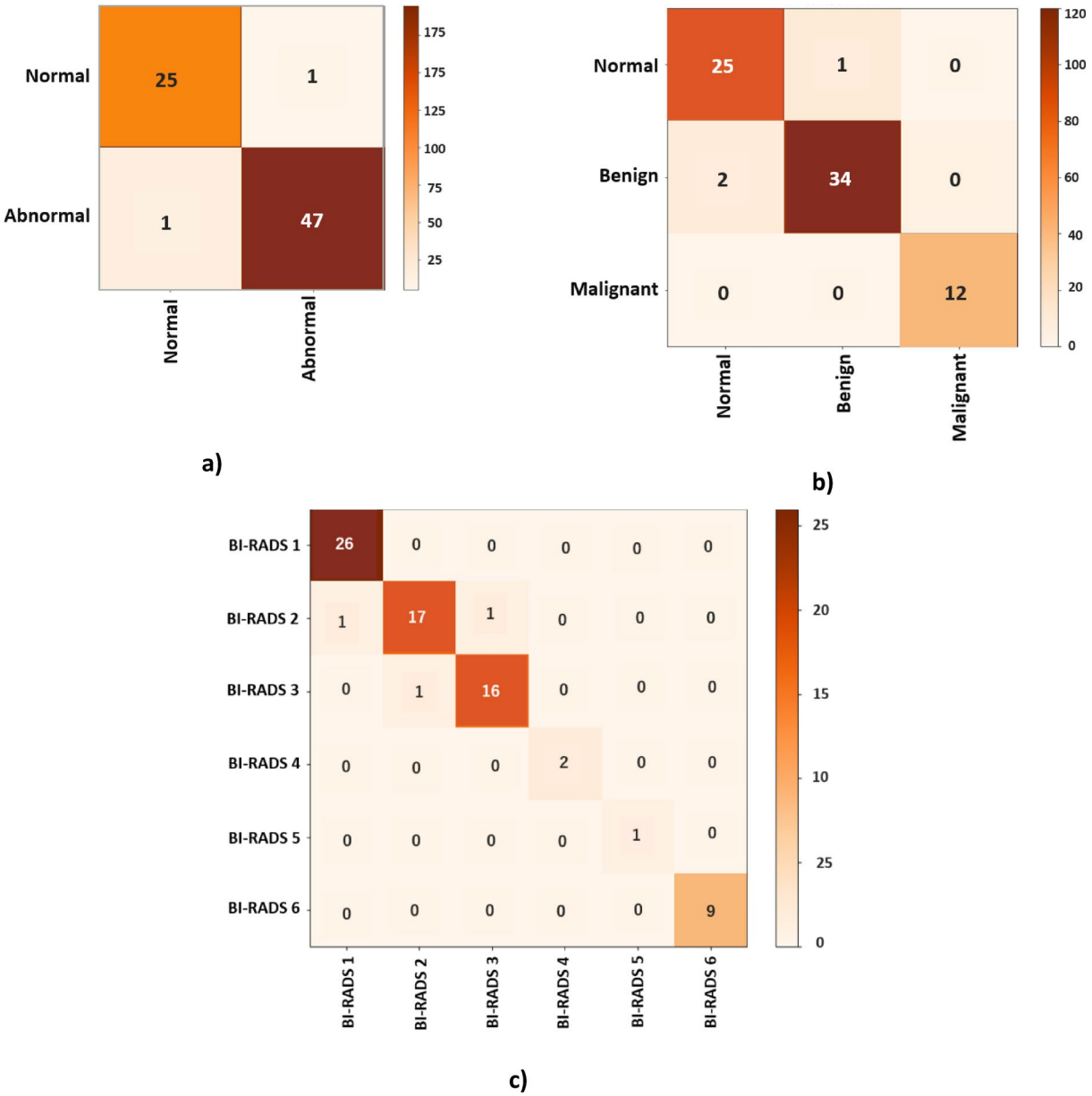

Fig. 6. The confusion matrices of the proposed CEET-Fed model in the centralized scenario: (**a**) binary task, (**b**) multi-class task, and (**c**) the BaggingClassifier model in BI-RADS task classification.

BI-RAD2 classes, which are the actual values. The probability ratios for abnormal and benign are 0.81 and 0.80, respectively. Nevertheless, the values were 0.45 and 0.41 for BI-RADS-2 and BI-RADS-3, respectively. Furthermore, the selection of features varies per approach, and the assignment of weights to each feature also differs. Subsequently, each feature was assigned a specific color code in the feature-value table to represent its association with the predicted abnormal (orange), not normal (blue) values.

Figure 9a presents the binary classification task that predicts whether a patient falls under a normal or abnormal category based on our dataset of health records' medical features. Axillary mass and related features are well-known clinical indicators of potential malignancy, making their strong influence on the prediction logical and medically interpretable. Furthermore, LIME allows the model's decision to be transparent by highlighting how less intuitive features (e.g., chewing qat) contribute to predictions, even if their medical relevance needs to be investigated further. Finally, the balance between "normal" and "abnormal" prediction probabilities underscores the robustness of the CEET-Fed model in weighting medically critical features.

Figure 9b presents the multi-class task that involves distinguishing between benign, normal, and malignant cases, which is crucial for stratifying patients based on the severity of risk. The reliance on well-documented clinical features such as shape or size change and regular menstruation demonstrates that the model aligns with medical intuition. By explaining the benign classification, LIME supports clinicians in understanding why a particular case was categorized as "Benign" and not as "Malignant." This can aid in reducing unnecessary

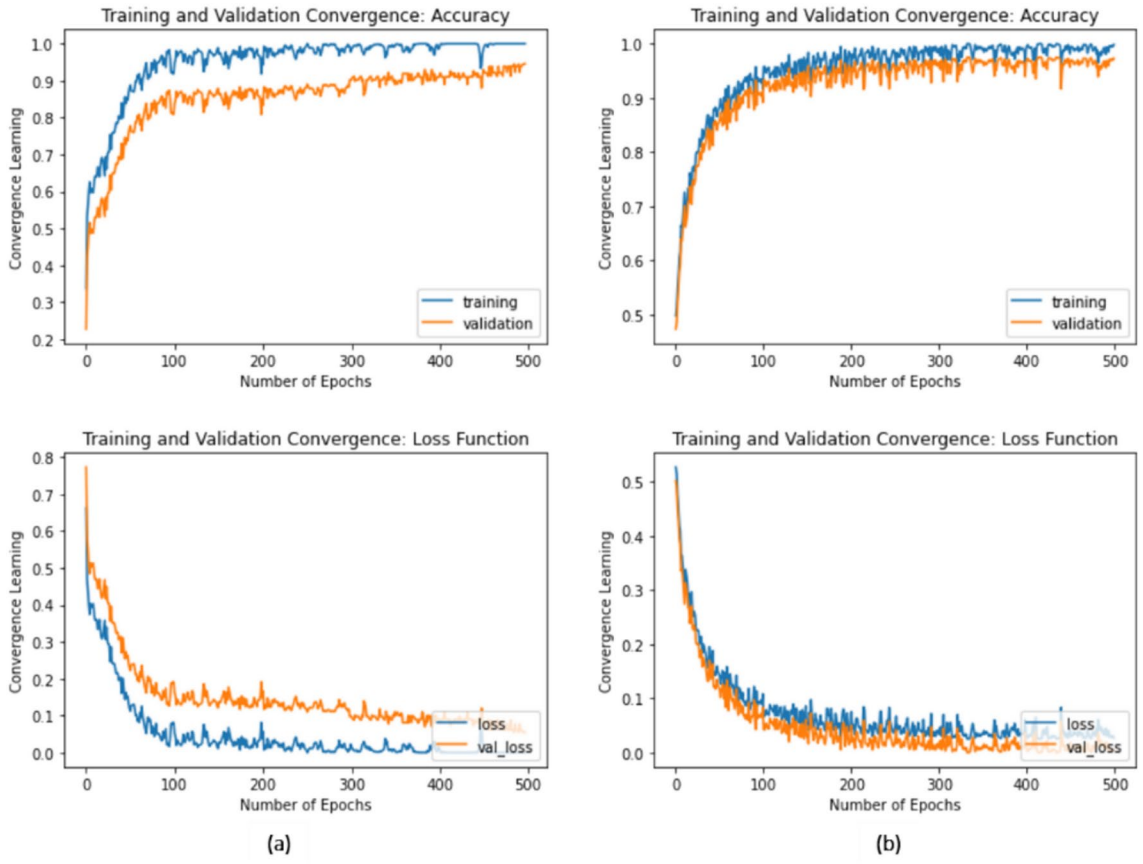

Fig. 7. The learning convergence rates of the best AI individual model and the proposed hybrid model through the binary task: (**a**) the Resnet50 model, (**b**) the proposed hybrid model.

Strategy	Models	Binary classification	Multi-class classification	BI-RADS classification
Individual model	LightGBM	0.900 [0.814–0.985]	0.916 [0.840–0.991]	0.937 [0.875–0.998]
	GradientBoosting	0.853 [0.748–0.956]	0.887 [0.803–0.970]	0.946 [0.888–1]
	XGBoost	0.872 [0.774–0.968]	0.928 [0.857–0.998]	0.942 [0.878–1]
	Catboost	0.881 [0.800–0.981]	0.923 [0.852–0.993]	0.964 [0.915–1]
	Random Forest	0.893 [0.798–0.987]	0.942 [0.878–1]	0.955 [0.901–1]
Late ensemble learning	Voting Ensemble	0.902 [0.813–0.990]	0.921 [0.848–0.993]	0.964 [0.915–1]
	BaggingClassifier (RF)	0.912 [0.825–0.997]	0.950 [0.888–1]	0.971 [0.926–1]
Individual model	VGG16	0.893 [0.797–0.987]	0.961 [0.909–1]	0.945 [0.920–1]
	ResNet 50	0.923 [0.863–0.982]	0.934 [0.873–0.994]	0.947 [0.892–1])
	Xception	0.865 [0.755–0.974]	0.883 [0.799–0.966]	0.907 [0.821–0.984]
Ensemble AI	Ensemble learning	0.942 [0.891–0.992]	0.956 [0.900–1]	0.970 [0.929–1]
Hybrid AI	Proposed hybrid CEET-Fed: Ensemble + ViT	0.970 [0.917–1]	0.956 [0.900–1]	0.970 [0.929–1]

Table 9. Evaluation of the AUC with 95% CI for all AI approaches and classification tasks (binary, multi-class, and BI-RADS classifications).

anxiety or overtreatment for patients. For features like BMI and shisha use, the explanations encourage further exploration into how these variables influence benign predictions, potentially leading to new insights into breast cancer risk factors.

Figure 9c presents the BI-RADS classification task, which categorizes breast imaging findings into standardized BI-RADS scores (e.g., Bi2, Bi3, Bi4) to stratify cancer likelihood and guide diagnostic steps. The presence of mass as a key feature aligns with established BI-RADS guidelines, where specific mass characteristics heavily influence category assignment. LIME's ability to highlight individual features, like menarche age or weight, underscores

Strategy	Models	Time complexity (seconds)
Individual ML model	LightGBM	0.17
	GradientBoosting	4.5
	XGBoost	0.40
	Catboost	0.45
	Random Forest	0.48
Late ensemble learning	Voting Ensemble	1.34
	BaggingClassifier (RF)	17.19
Individual DL model	VGG16	262.40
	ResNet 50	460.57
	Xception	240.20
Ensemble AI	Ensemble learning	474.24
Hybrid AI	Proposed hybrid CEET-Fed: Ensemble + ViT	592.24

Table 10. Comparison of time complexity across machine learning, deep learning, and hybrid ai strategies: balancing efficiency and computational demands.

Strategy	ACC	SEN	PRE	F1-score	AUC
CEET-Fed model in binary approach	98.65	98.65	98.68	98.64	98.10
CEET-Fed model in 3-class	97.30	97.30	97.37	97.27	98.10
CEET-Fed model in BI-RADS	95.95	95.95	95.62	95.90	97.17

Table 11. The evaluation results of the CEET-Fed model in federated learning scenario for all three approaches: Binary, 3-classes, and BI-RADS classification.

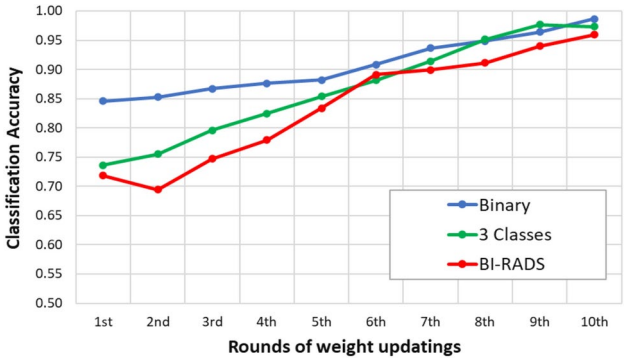

Fig. 8. The level of accuracy of the CEET-Fed model per round of updated weights to the server's global model in federated learning scenario for binary task, multi-class task, and BI-RADS classification task.

how additional non-imaging clinical features might complement traditional BI-RADS classification, enhancing its diagnostic utility.

Ablation study

The ablation study evaluates various machine learning strategies for breast cancer risk prediction using a dataset comprising clinical and biochemical features from 166 participants[27]. Eight features, including age, BMI, serum glucose levels, insulin levels, HOMA, leptin, adiponectin, resistin, and classification, were used to train and test the models. The study compares the performance of established ML methods, ensemble approaches, and the proposed CEET-Fed model under centralized and federated learning scenarios. The dataset is divided into 92 cases for training and validation and 24 cases for testing.

The dataset's clinical and biochemical features provide significant insights into breast cancer risk, as evidenced by the improved performance of advanced models like CEET-Fed. Features such as BMI, glucose levels, and adiponectin appear to play critical roles in model predictions, reinforcing the importance of incorporating diverse data sources for robust risk assessment. Figure 10 visualizes the heatmap correlation matrix for the features in a dataset related to breast cancer risk prediction.

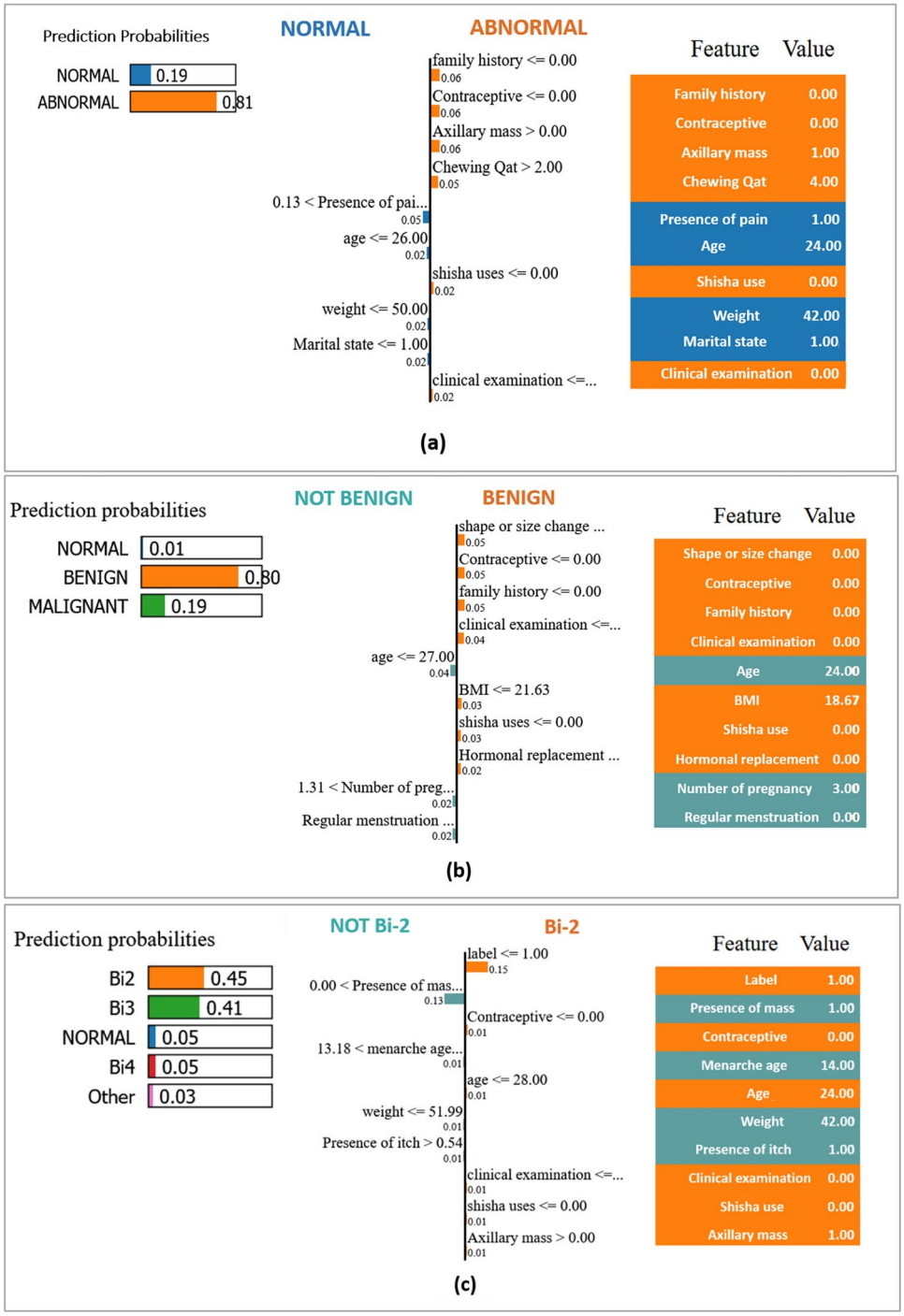

Fig. 9. LIME observations to explain the predictions of the proposed CEET-Fed model for all three tasks (**a**) the Binary task, (**b**) multi-class task, and (**c**) the BI-RADS task classification.

The CEET-Fed model in centralized learning achieved the highest performance among all strategies as shown in Table 12, with an ACC of 87.50%, SEN of 87.51%, PRE of 90.00%, F1-Score of 87.30%, and an AUC of 87.52%. Its 95% CI (71.58% to 98.61%) underscores the robustness and reliability of its predictions, outperforming traditional ML models and ensemble methods. The CEET-Fed model in federated learning demonstrated comparable performance to its centralized counterpart, achieving an ACC of 87.50%, SEN of 87.51%, and an AUC of 87.50%. While its CI (64.51% to 96.10%) is slightly narrower than the centralized model, it maintains reliable predictions. This result highlights the model's ability to perform effectively in distributed settings, preserving data privacy without compromising accuracy.

The federated learning implementation of CEET-Fed performed nearly as well as its centralized counterpart. This demonstrates the model's robustness in handling data distributed across multiple devices, preserving privacy

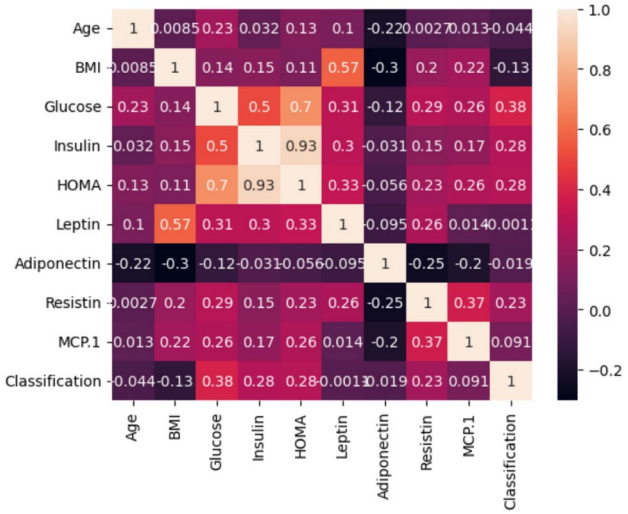

Fig. 10. Correlation heatmap of clinical and biochemical features used for breast cancer risk prediction.

Strategy	ACC	SEN	PRE	F1-score	AUC	95% CI
XGBoost	79.17	79.18	79.73	79.13	79.16	60.57% to 94.29%
Catboost	79.17	79.18	79.73	79.13	79.16	55.64% to 91.63%
Random Forest	83.33	83.34	83.33	83.31	83.35	62.62% to 95.26%
Voting Ensemble	83.33	83.34	83.33	83.31	83.35	62.62% to 95.26%
BaggingClassifier	79.17	79.17	81.11	78.84	79.17	52.34% to 89.64%
Proposed CEET-Fed model in centralized learning	87.50	87.51	90.00	87.30	87.52	71.58% to 98.61%
CEET-Fed model in federated learning	87.50	87.51	87.76	87.48	87.50	64.51% to 96.10%

Table 12. Performance evaluation of various machine learning strategies for breast cancer risk prediction using clinical and biochemical datasets.

while maintaining high predictive accuracy. The marginal difference in performance between centralized and federated learning settings underscores the feasibility of deploying CEET-Fed in real-world scenarios where data sharing is restricted. The ablation study highlights the effectiveness of the proposed CEET-Fed model for breast cancer risk prediction, showing its superiority over traditional ML and ensemble methods, even when dealing with heterogeneous data. The model's ability to deliver high accuracy in both centralized and federated learning scenarios demonstrates its potential for real-world healthcare applications, particularly in privacy-sensitive environments. The integration of clinical and biochemical features and the inclusion of robust statistical validation further establish CEET-Fed as a reliable and innovative solution for breast cancer prediction.

Figure 11 illustrates the prediction probabilities for classifying a participant as either "normal" or "abnormal" based on the CEET-Fed model. The predicted probability for the "Abnormal" class is 96%, indicating a strong likelihood of breast cancer risk. The LIME explanation visualizes the contribution of individual features to the prediction. Features such as resistin, glucose, HOMA, BMI, and MCP.1 positively influence the "abnormal" classification, while factors like age, adiponectin, and insulin slightly favor the "normal" classification. The feature importance chart on the right specifies the exact values for each feature, highlighting their contribution to the final decision.

Evaluation results of comparisons concerning breast cancer risk factors data and health records

This section presents a comparative evaluation of previous studies and our proposed model, highlighting their methodologies, datasets, and performance in predicting breast cancer risk based on health records and risk factor data. A summary of the findings is provided in the Table 13, showcasing the advancements in predictive accuracy over time. Previous works utilized various data sources, including linked health records, mammograms, electronic medical records, and biomarker clinical features, employing different methods of ML and DL and measuring using the AUC with 95% CI. Our proposed model, combining CNN with ViT and employing federated learning, demonstrated superior performance, achieving an AUC of 0.970 (95% CI: 0.929–1) on a dataset of 734 participants. By integrating risk factors from electronic health records (EHR) and symptoms, our model significantly outperforms existing approaches, showcasing the potential of advanced deep learning techniques and federated learning in breast cancer risk prediction. These results emphasize the importance of leveraging diverse data sources and state-of-the-art algorithms for improved predictive accuracy.

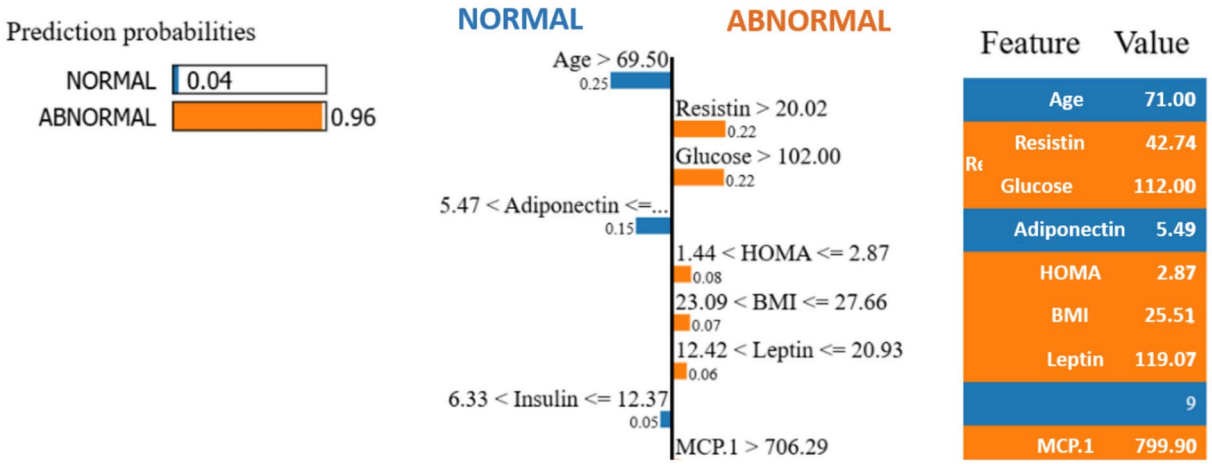

Fig. 11. LIME-based explanation of the CEET-Fed model's prediction for breast cancer risk classification.

Pervious study	Dataset	Number of patients	Methods	AUC with 95% CI
A. Akselrod-Ballin et al.[25]	Linked health records and mammograms	Validated in 1055 women	DNN, XGboost	(AUC) of 0.91 (95% CI: 0.89, 0.93)
A. Yala et al.[26]	Patient questionnaires and electronic medical records review	Test set included 3937 women	Risk-factor-based logistic regression model (RF-LR)	(AUC) of 0.67 (95% CI: 0.62, 0.72)
M. Patrício et al.[27]	Biomarker clinical features	166 participants	SVM	95% CI: [0.87, 0.91]
A. Michel et al.[28]	Electronic health record (EHR) data on risk factors	23,467 women	CNN	(AUC 0.845 vs. 0.589)
The proposed model	Risk factors from HER and symptoms	734 participants	CNN + ViT with federated learning	(AUC) of 0.970 (95% CI: 0.929–1)

Table 13. Comparative evaluation of breast cancer risk prediction models based on health records and risk factor data.

Limitations and future works

Gathering data, especially medical data, is a costly and laborious operation, requiring significant work, time, and financial resources in addition to the use of a dedicated medical staff to facilitate the process of collecting and understanding the data. In this research, the risk factors of breast cancer and patient medical records were collected along with radiological images and medical reports for 734 cases, while AI models require a large amount of data. For interpretation, ambiguity in borderline cases, such as Bi2 versus Bi3, limits prediction accuracy. LIME explanations highlight these challenges, suggesting the need for further model refinement. AI models were conducted on the risk factor dataset while the radiologists completed processing and identifying the region of interest in the image data for each patient. So, a hybrid model might be applied to the mammogram and ultrasound images[25] with the breast cancer risk factor data set as future work. Furthermore, provides an end-to-end system for segmentation and detection utilizing YOLO versions[80] and AI report generation[81]. Additionally, we propose future work exploring model compression techniques and the deployment of edge computing solutions to make the model more accessible in such settings.

Conclusion

This work involved the implementation of both a centralized learning scenario without federated and a federated learning scenario to predict the occurrence of breast cancer using risk factors and patients' medical records. Data was gathered, and risk factors were identified based on health records and medical reports regarding mammograms or ultrasound images. Radiologists generated the reports, which were then classified according to the BI-RADS scores (1–6) as either benign, malignant, or normal. In the centralized scenario, machine learning models were utilized alone, and subsequently, ensemble learning was employed to combine the most optimal outcomes with high accuracy. In addition, we separately utilized deep learning and transfer learning models. Furthermore, we proposed a CEET-Fed model that implemented ensemble learning with a transformer encoder model. All models are carried out using three different approaches: binary classification, three-class classification, and BI-RADS classification. The proposed CEET-Fed model attained an accuracy of 97.30%, with binary and three-class accuracy levels of 95.95% and three-class, respectively. The BaggingClassifier achieved the highest accuracy of 95.95% in the BI-RADS classification approach to explaining the predictions. In addition, the proposed CEET-Fed implements a federated learning setting that keeps each client's privacy regarding breast risk data while addressing the issue of imbalanced data. Compared to the centralized scenario, the federated

model showed superior performance, achieving accuracy rates of 98.65% for the binary classification and 97.30% for the three-class classification approach, with XAI used to explain the predictions.

Data availability

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Received: 13 October 2024; Accepted: 28 March 2025
Published online: 27 May 2025

References

1. A. C. Society and L. Department. American Cancer Society: 'Breast Cancer Facts & Figures 2022–2024. https://www.breastcancer.org/facts-statistics. (2023).
2. Breast cancer. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/breast-cancer
3. Houssein, E. H. et al. Using deep DenseNet with cyclical learning rate to classify leukocytes for leukemia identification. *Front. Oncol.* https://doi.org/10.3389/fonc.2023.1230434 (2023).
4. Beltran-Bless, A.-A. et al. Regularly scheduled physical examinations and the detection of breast cancer recurrences. *Breast* **69**, 274–280. https://doi.org/10.1016/j.breast.2023.03.004 (2023).
5. Muduli, D., Dash, R. & Majhi, B. Automated diagnosis of breast cancer using multi-modal datasets: A deep convolution neural network based approach. *Biomed. Signal Process. Control* **71**, 102825. https://doi.org/10.1016/j.bspc.2021.102825 (2022).
6. Hamedani-KarAzmoudehFar, F., Tavakkoli-Moghaddam, R., Tajally, A. R. & Aria, S. S. Breast cancer classification by a new approach to assessing deep neural network-based uncertainty quantification methods. *Biomed. Signal Process. Control* **79**, 104057. https://doi.org/10.1016/j.bspc.2022.104057 (2023).
7. Abdar, M. & Makarenkov, V. CWV-BANN-SVM ensemble learning classifier for an accurate diagnosis of breast cancer. *Measurement* **146**, 557–570. https://doi.org/10.1016/j.measurement.2019.05.022 (2019).
8. Sahu, A., Das, P. K. & Meher, S. An efficient deep learning scheme to detect breast cancer using mammogram and ultrasound breast images. *Biomed. Signal Process. Control* **87**, 105377. https://doi.org/10.1016/j.bspc.2023.105377 (2024).
9. W Wolberg. Breast Cancer Wisconsin (Original). *UCI Mach. Learn. Repos.* https://doi.org/10.24432/C5HP4Z. (1992).
10. Al-Tam, R. M. et al. A hybrid framework of transformer encoder and residential conventional for cardiovascular disease recognition using heart sounds. *IEEE Access* **12**, 123099–123113. https://doi.org/10.1109/ACCESS.2024.3451660 (2024).
11. Myrzashova, R., Alsamhi, S. H., Shvetsov, A. V., Hawbani, A. & Wei, X. Blockchain meets federated learning in healthcare: A systematic review with challenges and opportunities. *IEEE Internet Things J.* **10**(16), 14418–14437. https://doi.org/10.1109/JIOT.2023.3263598 (2023).
12. Sharma, S. & Guleria, K. A comprehensive review on federated learning based models for healthcare applications. *Artif. Intell. Med.* **146**, 102691. https://doi.org/10.1016/j.artmed.2023.102691 (2023).
13. Li, H. et al. Review on security of federated learning and its application in healthcare. *Futur. Gener. Comput. Syst.* **144**, 271–290. https://doi.org/10.1016/j.future.2023.02.021 (2023).
14. Antoniadi, A. M. et al. Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: A systematic review. *Appl. Sci.* **11**(11), 5088. https://doi.org/10.3390/app11115088 (2021).
15. Nasarian, E., Alizadehsani, R., Acharya, U. R. & Tsui, K.-L. Designing interpretable ML system to enhance trust in healthcare: A systematic review to proposed responsible clinician-AI-collaboration framework. *Inf. Fusion* **108**, 102412. https://doi.org/10.1016/j.inffus.2024.102412 (2024).
16. Di Martino, F. & Delmastro, F. Explainable AI for clinical and remote health applications: a survey on tabular and time series data. *Artif. Intell. Rev.* **56**(6), 5261–5315. https://doi.org/10.1007/s10462-022-10304-3 (2023).
17. Demir, S. & Sahin, E. K. Predicting occurrence of liquefaction-induced lateral spreading using gradient boosting algorithms integrated with particle swarm optimization: PSO-XGBoost, PSO-LightGBM, and PSO-CatBoost. *Acta Geotech.* **18**(6), 3403–3419. https://doi.org/10.1007/s11440-022-01777-1 (2023).
18. Chen, T. & Guestrin, C. XGBoost. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (ACM, 2016).
19. A Prokhorenkova, L Gusev, G Vorobev, A Dorogush, AV Gulin. CatBoost: unbiased boosting with categorical features. *Adv. Neural Inf. Process. Syst.* **31**, 6638–6648 (2018).
20. Biau, G. & Scornet, E. A random forest guided tour. *TEST* **25**(2), 197–227. https://doi.org/10.1007/s11749-016-0481-7 (2016).
21. Alkhpor, H. K. & Alserhani, F. M. Collaborative federated learning-based model for alert correlation and attack scenario recognition. *Electronics* **12**(21), 4509. https://doi.org/10.3390/electronics12214509 (2023).
22. Łukasiewicz, S. et al. Breast cancer—Epidemiology, risk factors, classification, prognostic markers, and current treatment strategies—An updated review. *Cancers* **13**(17), 4287. https://doi.org/10.3390/cancers13174287 (2021).
23. Roberts, E., Howell, S. & Evans, D. G. Polygenic risk scores and breast cancer risk prediction. *Breast* **67**, 71–77. https://doi.org/10.1016/j.breast.2023.01.003 (2023).
24. Britt, K. L., Cuzick, J. & Phillips, K.-A. Key steps for effective breast cancer prevention. *Nat. Rev. Cancer* **20**(8), 417–436. https://doi.org/10.1038/s41568-020-0266-x (2020).
25. Akselrod-Ballin, A. et al. Predicting breast cancer by applying deep learning to linked health records and mammograms. *Radiology* **292**(2), 331–342. https://doi.org/10.1148/radiol.2019182622 (2019).
26. Yala, A., Lehman, C., Schuster, T., Portnoi, T. & Barzilay, R. A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology* **292**(1), 60–66. https://doi.org/10.1148/radiol.2019182716 (2019).
27. Patrício, M. et al. Using resistin, glucose, age and BMI to predict the presence of breast cancer. *BMC Cancer* **18**(1), 29. https://doi.org/10.1186/s12885-017-3877-1 (2018).
28. Michel, A. et al. Breast cancer risk prediction combining a convolutional neural network-based mammographic evaluation with clinical factors. *Breast Cancer Res. Treat.* **200**(2), 237–245. https://doi.org/10.1007/s10549-023-06966-4 (2023).
29. Wolberg, W., Mangasarian, W., Street, O. & Street, N. Breast cancer Wisconsin (diagnostic). *UCI Mach. Learn. Repos.* https://doi.org/10.24432/C5DW2B (1995).
30. Alshayeji, M. H., Ellethy, H., Abed, S. & Gupta, R. Computer-aided detection of breast cancer on the Wisconsin dataset: An artificial neural networks approach. *Biomed. Signal Process. Control* **71**, 103141. https://doi.org/10.1016/j.bspc.2021.103141 (2022).
31. Ali, M. S., Hossain, M. M., Kona, M. A., Nowrin, K. R. & Islam, M. K. An ensemble classification approach for cervical cancer prediction using behavioral risk factors. *Healthc. Anal.* **5**, 100324. https://doi.org/10.1016/j.health.2024.100324 (2024).
32. Zhao, Z., Alzubaidi, L., Zhang, J., Duan, Y. & Gu, Y. A comparison review of transfer learning and self-supervised learning: Definitions, applications, advantages and limitations. *Expert Syst. Appl.* **242**, 122807. https://doi.org/10.1016/j.eswa.2023.122807 (2024).

33. Mazni, M. et al. An investigation into real-time surface crack classification and measurement for structural health monitoring using transfer learning convolutional neural networks and Otsu method. *Alex. Eng. J.* **92**, 310–320. https://doi.org/10.1016/j.aej.2024.02.052 (2024).
34. Addo, D. et al. A hybrid lightweight breast cancer classification framework using the histopathological images. *Biocybern. Biomed. Eng.* **44**(1), 31–54. https://doi.org/10.1016/j.bbe.2023.12.003 (2024).
35. Senan, E. A., Alsaade, F. W., Al-Mashhadani, M. I. A. & Theyazn, H. H. Classification of histopathological images for early detection of breast cancer using deep learning". *J. Appl. Sci. Eng.* **24**(3), 323–329 (2021).
36. Shamsan, A., Senan, E. M. & Ahmad Shatnawi, H. S. Predicting of diabetic retinopathy development stages of fundus images using deep learning based on combined features. *PLoS ONE* **18**, e0289555. https://doi.org/10.1371/journal.pone.0289555 (2023).
37. Prusty, S., Patnaik, S., Dash, S. K. & Priyadarsini Prusty, S. G. SEMeL-LR: An improvised modeling approach using a meta-learning algorithm to classify breast cancer". *Eng. Appl. Artif. Intell.* **129**(1), 107630. https://doi.org/10.1016/j.engappai.2023.107630 (2024).
38. Al-Hejri, A. M. et al. ETECADx: Ensemble self-attention transformer encoder for breast cancer diagnosis using full-field digital X-ray breast images. *Diagnostics* **13**(1), 89. https://doi.org/10.3390/diagnostics13010089 (2022).
39. Ukwuoma, C. C. et al. A hybrid explainable ensemble transformer encoder for pneumonia identification from chest X-ray images. *J. Adv. Res.* **48**, 191–211. https://doi.org/10.1016/j.jare.2022.08.021 (2023).
40. Al-Tam, R. M. et al. A hybrid workflow of residual convolutional transformer encoder for breast cancer classification using digital X-ray mammograms. *Biomedicines* **10**(11), 2971. https://doi.org/10.3390/biomedicines10112971 (2022).
41. Abdar, M. et al. A new nested ensemble technique for automated diagnosis of breast cancer. *Pattern Recognit. Lett.* **132**, 123–131. https://doi.org/10.1016/j.patrec.2018.11.004 (2020).
42. Zhou, G. & Mosadegh, B. Distilling knowledge from an ensemble of vision transformers for improved classification of breast ultrasound. *Acad. Radiol.* https://doi.org/10.1016/j.acra.2023.08.006 (2023).
43. Springenberg, M. et al. From modern CNNs to vision transformers: Assessing the performance, robustness, and classification strategies of deep learning models in histopathology. *Med. Image Anal.* **87**, 102809. https://doi.org/10.1016/j.media.2023.102809 (2023).
44. Baccouche, A., Garcia-Zapirain, B. & Elmaghraby, A. S. An integrated framework for breast mass classification and diagnosis using stacked ensemble of residual neural networks. *Sci. Rep.* **12**(1), 12259. https://doi.org/10.1038/s41598-022-15632-6 (2022).
45. He, Q., Yang, Q., Su, H. & Wang, Y. Multi-task learning for segmentation and classification of breast tumors from ultrasound images. *Comput. Biol. Med.* **173**, 108319. https://doi.org/10.1016/j.compbiomed.2024.108319 (2024).
46. Bukhari, S. M. S. et al. Secure and privacy-preserving intrusion detection in wireless sensor networks: Federated learning with SCNN-Bi-LSTM for enhanced reliability. *Ad Hoc Netw.* **155**, 103407. https://doi.org/10.1016/j.adhoc.2024.103407 (2024).
47. Hasan, M. K. et al. Federated learning enables 6 G communication technology: Requirements, applications, and integrated with intelligence framework. *Alex. Eng. J.* **91**, 658–668. https://doi.org/10.1016/j.aej.2024.02.044 (2024).
48. Zhao, Z. et al. Towards efficient communications in federated learning: A contemporary survey. *J. Franklin Inst.* **360**(12), 8669–8703. https://doi.org/10.1016/j.jfranklin.2022.12.053 (2023).
49. Banabilah, S., Aloqaily, M., Alsayed, E., Malik, N. & Jararweh, Y. Federated learning review: Fundamentals, enabling technologies, and future applications. *Inf. Process. Manag.* **59**(6), 103061. https://doi.org/10.1016/j.ipm.2022.103061 (2022).
50. Zhao, L., Xie, H., Zhong, L. & Wang, Y. Explainable federated learning scheme for secure healthcare data sharing. *Heal. Inf. Sci. Syst.* **12**(1), 49. https://doi.org/10.1007/s13755-024-00306-6 (2024).
51. Wang, W. et al. A privacy preserving framework for federated learning in smart healthcare systems. *Inf. Process. Manag.* **60**(1), 103167. https://doi.org/10.1016/j.ipm.2022.103167 (2023).
52. Almufareh, M. F., Tariq, N., Humayun, M. & Almas, B. A federated learning approach to breast cancer prediction in a collaborative learning framework. *Healthcare* **11**(24), 3185. https://doi.org/10.3390/healthcare11243185 (2023).
53. Peta, J. & Koppu, S. Enhancing breast cancer classification in histopathological images through federated learning framework. *IEEE Access* **11**, 61866–61880. https://doi.org/10.1109/ACCESS.2023.3283930 (2023).
54. Truhn, D. et al. Encrypted federated learning for secure decentralized collaboration in cancer image analysis. *Med. Image Anal.* **92**, 103059. https://doi.org/10.1016/j.media.2023.103059 (2024).
55. Nicholson, B. T. et al. Accuracy of assigned BI-RADS breast density category definitions. *Acad. Radiol.* **13**(9), 1143–1149. https://doi.org/10.1016/j.acra.2006.06.005 (2006).
56. Galli, S. *Python Feature Engineering Cookbook: Over 70 Recipes for Creating, Engineering, and Transforming Features to Build Machine Learning Models* (Packt Publishing Ltd, 2022).
57. Al-antari, M. A. et al. An automatic computer-aided diagnosis system for breast cancer in digital mammograms via deep belief network. *J. Med. Biol. Eng.* **38**(3), 443–456. https://doi.org/10.1007/s40846-017-0321-6 (2018).
58. Zebari, R., Abdulazeez, A., Zeebaree, D., Zebari, D. & Saeed, J. A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction". *J. Appl. Sci. Technol. Trends* **1**, 56–70 (2020).
59. Nguyen, C., Wang, Y. & Nguyen, H. N. Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. *J. Biomed. Sci. Eng.* **06**(05), 551–560. https://doi.org/10.4236/jbise.2013.65070 (2013).
60. Nguyen, D. C. et al. Federated learning for smart healthcare: A survey. *ACM Comput. Surv.* **55**(3), 1–37. https://doi.org/10.1145/3501296 (2023).
61. Ferrag, M. A., Friha, O., Hamouda, D., Maglaras, L. & Janicke, H. Edge-IIoTset: A new comprehensive realistic cyber security dataset of IoT and IIoT applications for centralized and federated learning. *IEEE Access* **10**, 40281–40306. https://doi.org/10.1109/ACCESS.2022.3165809 (2022).
62. Fernandez, A., Garcia, S., Herrera, F. & Chawla, N. V. SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *J. Artif. Intell. Res.* **61**, 863–905. https://doi.org/10.1613/jair.1.11192 (2018).
63. Fauzi, M. A., Yang, B. & Blobel, B. Comparative analysis between individual, centralized, and federated learning for smartwatch based stress detection. *J. Pers. Med.* **12**(10), 1584. https://doi.org/10.3390/jpm12101584 (2022).
64. Chandramouli, A., Hyma, V. R., Tanmayi, P. S., Santoshi, T. G. & Priyanka, B. Diabetes prediction using hybrid bagging classifier. *Entertain. Comput.* **47**, 10059. https://doi.org/10.1016/j.entcom.2023.100593 (2023).
65. Khan, M. A., Iqbal, N., Jamil, H. & Kim, D. H. An optimized ensemble prediction model using AutoML based on soft voting classifier for network intrusion detection. *J. Netw. Comput. Appl.* **212**, 103560. https://doi.org/10.1016/j.jnca.2022.103560 (2023).
66. S Mascarenhas, M Agarwal. A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for Image Classification. In *2021 International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications (CENTCON)*. IEEE, 96–99. https://doi.org/10.1109/CENTCON52345.2021.9687944. (2021).
67. TD Long, JE Shelhamer. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. (2015).
68. Shaheed, K. et al. DS-CNN: A pre-trained Xception model based on depth-wise separable convolutional neural network for finger vein recognition. *Expert Syst. Appl.* **191**, 116288. https://doi.org/10.1016/j.eswa.2021.116288 (2022).
69. Haq, I. U., Ali, H., Wang, H. Y., Lei, C. & Ali, H. Feature fusion and ensemble learning-based CNN model for mammographic image classification. *J. King Saud. Univ. Comput. Inf. Sci.* **34**(6), 3310–3318. https://doi.org/10.1016/j.jksuci.2022.03.023 (2022).
70. Aurna, N. F., Yousuf, M. A., Taher, K. A., Azad, A. K. M. & Moni, M. A. A classification of MRI brain tumor based on two stage feature level ensemble of deep CNN models. *Comput. Biol. Med.* **146**, 105539. https://doi.org/10.1016/j.compbiomed.2022.105539 (2022).
71. Han, Y. et al. Transformer in transformer. *Adv. Neural Inf. Process. Syst.* **34**, 15908–15919 (2021).

72. Vaswani, A. et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, 1 (2017).
73. Jiménez-Sánchez, A., Tardy, M., González Ballester, M. A., Mateus, D. & Piella, G. Memory-aware curriculum federated learning for breast cancer classification. *Comput. Methods Progr. Biomed.* **229**, 107318. https://doi.org/10.1016/j.cmpb.2022.107318 (2023).
74. McMahan, BA., Moore, E., Ramage, D., Hampson, S. & Arcas, BA. Communication-efficient learning of deep networks from decentralized data. In Artificial Intelligence and Statistics, 1273–1282 (PMLR, 2017).
75. Mitchell, J., McLaughlin, N. & Martinez-del-Rincon, J. Generating sparse explanations for malicious Android opcode sequences using hierarchical LIME. *Comput. Secur.* **137**, 103637. https://doi.org/10.1016/j.cose.2023.103637 (2024).
76. Ribeiro, M. T., Singh, S. & Guestrin, C. Why Should I Trust You? In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ed. Kani, B.) 1135–1144 (ACM, 2016).
77. Patil, S. et al. Explainable artificial intelligence for intrusion detection System. *Electronics* **11**(19), 3079. https://doi.org/10.3390/electronics11193079 (2022).
78. Li, C., Zhou, J., Du, K. & Dias, D. Stability prediction of hard rock pillar using support vector machine optimized by three metaheuristic algorithms. *Int. J. Min. Sci. Technol.* **33**(8), 1019–1036. https://doi.org/10.1016/j.ijmst.2023.06.001 (2023).
79. Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**(1), 29–36. https://doi.org/10.1148/radiology.143.1.7063747 (1982).
80. Al-Tam, R. M., Al-Hejri, A. M., Alshamrani, S. S., Al-antari, M. A. & Narangale, S. M. Multimodal breast cancer hybrid explainable computer-aided diagnosis using medical mammograms and ultrasound Images. *Biocybern. Biomed. Eng.* **44**(3), 731–758. https://doi.org/10.1016/j.bbe.2024.08.007 (2024).
81. Li, J., Tang, T., Zhao, W. X., Nie, J.-Y. & Wen, J.-R. Pre-trained language models for text generation: A survey. *ACM Comput. Surv.* **56**(9), 1–39. https://doi.org/10.1145/3649449 (2024).

Acknowledgements

We express our gratitude to the National Cancer Control Foundation (NCCF) in Sanaa, the Republic of Yemen represented by the head of the center Thaker Al-absi, and Dr. Asma Ameen for facilitating and supporting the team in collecting data and discussing it with doctors specializing in cancer and radiology.

Author contributions

Aymen M. Al-Hejri: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Resources; Software; Validation; Visualization; Writing – original draft. Riyadh M. Al-Tam: Conceptualization Data curation; Formal analysis; Investigation; Methodology; Resources; Validation; Writing – original draft. Mugahed A. Al-antari: Formal analysis; Investigation; Resources; Validation; Visualization; Writing – review & editing. Sultan S. Alshamrani: Formal analysis; Funding acquisition; Investigation; Project administration; Resources; Validation; Visualization; Writing – review & editing. Kaled M Alshmrany: Formal analysis; Funding acquisition; Investigation; Project administration; Resources; Validation; Visualization; Writing – review & editing. Archana Harsing Sable: Conceptualization; Methodology; Investigation; Data Curation; Formal Analysis; Writing—review & editing; Resources; Validation; Supervision. Wedad Alatebi: Methodology; Formal Analysis; Writing—review & editing; Validation

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.M.A.-H.

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.