ORIGINAL ARTICLE

MSAE-DL: enhancing breast cancer classification through hybrid self-attention integration, feature fusion, and ensemble classification in digital breast tomosynthesis

Alaa M. Adel El-Shazli[1] · **Sherin M. Youssef**[1] · **Abdel Hamid Soliman**[2] · **Claude Chibelushi**[3]

Received: 14 October 2024 / Accepted: 19 March 2025 / Published online: 29 May 2025

Abstract

Breast cancer is a significant global health challenge, with highest rates of occurrence and mortality worldwide. Early detection is important for the improvement of patient outcomes and the reduction of the overall burden of the disease. Digital breast tomosynthesis (DBT) scans offer three-dimensional images of the breast tissue and is becoming a valuable tool in the detection of breast abnormalities. However, accurately classifying DBT scans is challenging due to the complexity of the anatomy of the breast and the presence of minor abnormalities. This study introduces the MSAE-DL system for the multi-class classification of DBT scans. The system incorporates a novel multi-head self-attention model with a unique ensemble classification model. Features were extracted from the Mod_AlexNet Self-Attention model and fused with histogram of oriented gradients (HOG) descriptors. Subsequently, feature vectors are reduced using three feature selection models. Finally, a novel ensemble classification model is introduced and fuses class and classifier weights for the final prediction using various classifiers. The system demonstrates optimal performance in classifying DBT scans into normal, benign, and malignant classes, achieving an accuracy of 90.13%, precision of 92.77%, and f1-score of 91.03%. The experimental results underscore the potential of this approach in enhancing DBT classification into three different classes, rather than simply binary classification.

Keywords Digital breast tomosynthesis · Classification · Ensemble classification · Self-attention

1 Introduction

The International Agency for Research on Cancer (IARC) released updated estimates of the worldwide cancer incidence in 2020, indicating a rise of 19.3 million new cases and 10.0 million deaths [1]. Globally, one in five individuals will have cancer at some point in their lives, and one in eight men and one in eleven women will pass away from the disease. The most alarming finding from the updated IARC predictions is that, with 2.3 million new instances of female breast cancer identified in 2020—more than the number of new cases of lung cancer for the first time—breast cancer is now the most commonly diagnosed cancer globally. Currently, 11.7% of newly diagnosed cases of cancer in both sexes are breast cancer cases [1]. Performing a clinical breast examination every two years results in a 15% overall decline in breast cancer mortality, with a considerable reduction of almost 30%

https://doi.org/10.1007/s00521-025-11192-8

in mortality for women aged 50 and above [2]. This examination also considerably downstages breast cancer at diagnosis [2]. Therefore, it is critical to utilize cutting-edge breast screening technologies for early detection and diagnosis to lower mortality rates and end the worldwide burden of cancer. A significant advancement in breast imaging technology, Digital Breast Tomosynthesis (DBT), provides a three-dimensional image of the breast tissue. In contrast to conventional mammography, which produces a two-dimensional picture, DBT reconstructs a number of high-resolution slices from many X-ray images taken from various angles [3]. The technology addresses the concerns caused by overlapping breast tissues in traditional mammography, reducing false positives and false negatives, and improving cancer diagnosis.

Breast cancer detection and classification using deep learning algorithms have shown promising results in recent studies. The use of convolutional neural networks has enabled more accurate and efficient analysis of mammograms and other medical images for the early detection of breast cancer. These algorithms can assist in identifying patterns and anomalies that might not be visible to the human eye, leading to improved diagnostic accuracy and potential early intervention. By leveraging the power of deep learning algorithms, specifically convolutional networks, significant advancements have been made in the field of medical image analysis for breast cancer detection and classification [4].

Computer-based detection technologies have become crucial for enhancing the interpretation of Digital Breast Tomosynthesis imaging. DBT images provide three-dimensional details about the breast and automated detection methods are valuable for identifying small abnormalities such as lumps or calcifications. To aid radiologists with diagnosis, computer-aided diagnostic systems employ advanced algorithms leading to improved accuracy and efficiency in recognizing potential abnormalities. Numerous CAD systems have been developed to detect breast cancer in DBT scans. These systems face challenges such as the lack of multi-class classification data of DBT, challenges in differentiating between benign and malignant tumours, and handling variations in breast density and size that affect the precise classification by automated CAD systems. Most of the systems developed so far for DBT scan classification have focused on simple binary categories. They either classify scans as benign or malignant, cancerous versus non-cancerous (where non-cancerous includes both benign and normal cases), or normal versus abnormal (where abnormal covers both benign and malignant cases). While these methods serve a purpose, they fall short in addressing the more complex, but crucial, need for multi-class classification. By lumping different conditions together, these systems miss the nuances that could significantly improve early detection and diagnosis. Multi-class classification—separating benign, normal, and malignant scans—offers a more accurate and meaningful approach, helping to improve patient outcomes.

One research focus is continuously enhancing image quality using cutting-edge methods to address artefact reduction, contrast-to-noise ratio, and spatial resolution in DBT scans. Researchers like Gao et al., Gao, Fessler and Chan, Su et al., Siti Noraini Sulaiman et al., Syafiqah Aqilah Saifudin et al., Mota, Mendes, and Matela [5–10] have made significant contributions in this area of study. Reducing patient exposure while maintaining diagnostic precision was the aim of researchers studying radiation dose optimization including Ajay Kumar Visvkarma et al. [11], as demonstrated by their work. These investigations also thoroughly investigate the impact of different acquisition conditions on the quality of images. Several studies have been conducted on the automated classification of tomosynthesis scans, with the majority of research focusing on binary classification.

Chen et al. (2024) [12] introduced Deep-AutoMO, a multi-objective neural network designed to classify benign and malignant lesions in Digital Breast Tomosynthesis (DBT) images. The model combines two innovative techniques: Multi-objective Immune Neural Architecture Search (MINAS) and Evidential Reasoning based on Entropy (ERE). MINAS focuses on optimizing sensitivity and specificity during training by generating a set of deep neural networks (DNNs) that blend ResNet and DenseNet blocks with pooling layers, using Bayesian optimization to fine-tune their performance. This ensures a balance between sensitivity and specificity, addressing the common challenge of class imbalance in DBT data. ERE, applied during testing, enhances the robustness of the model by estimating prediction uncertainty and integrating outputs from multiple DNNs. This approach ensures reliable and accurate predictions, even in noisy or out-of-distribution scenarios. Deep-AutoMO achieved an accuracy of 85.57%, a specificity of 87.68%, and an AUC of 89.25%. Their work is limited by the use of a private dataset, which may affect the generalizability of the findings, and by focusing exclusively on binary classification of benign and malignant cases, leaving normal cases unaddressed.

Shao et al. (2024) [13] explored the use of AI-based techniques to classify small breast masses ($\leq$ 2 cm) using digital mammography (DM), Digital Breast Tomosynthesis (DBT), and a combination of both (DM + DBT). They developed two types of models: radiomics models, which rely on manually extracted image features, and deep learning models, which automatically learn features from the data using a ResNet-34 architecture. The combined DM + DBT models consistently delivered better performance compared to using DM or DBT alone, with the deep learning DM + DBT model achieving the highest AUC of 0.908 on the internal dataset. External validation further supported the benefits of combining DM and DBT, particularly for detecting small tumours. The study demonstrated that deep learning outperformed radiomics in accuracy and robustness, highlighting the potential of DBT to enhance breast cancer diagnosis. However, the work faced limitations, including a relatively small dataset, manual segmentation of tumour regions, and a focus solely on binary classification of benign versus malignant cases.

Oladimeji et al. (2024) [14] introduced an advanced framework called mutual information-based radiomic feature selection (MIRFS), combined with SHAP explainability, to help classify Digital Breast Tomosynthesis (DBT) scans into benign or malignant classes. Their study used a subset of the BCS-DBT dataset, focusing on 31 benign and 26 malignant cases, to test and validate the approach. By identifying the 15 most important radiomic features using mutual information, the framework tackled challenges like feature redundancy and captured complex patterns in the data, leading to improved accuracy. When applied with a Random Forest classifier, the system achieved outstanding results of 92% accuracy, 93% precision, and a 92% F1-score, outperforming traditional methods like LASSO and RFE, as well as deep learning models. Their study has some limitations, including the use of a relatively small dataset, the focus solely on binary classification without exploring multi-class scenarios.

Farangis Sajadi Moghadam and Rashidi (2024) [15] developed a feature extraction model using the Discrete Cosine-based Stockwell Transform (DCT-DOST), and radiomic features, to classify DBT images into benign or malignant. Their approach involved pre-processing, segmentation, feature extraction, and classification. Synthetic minority oversampling technique (SMOTE) was deployed for feature selection, while Random Forest (RF), K-Nearest Neighbour (KNN), and Support Vector Machine (SVM) were employed for the classification stage. The best results were achieved by the RF classifier with an accuracy of 78.51%, and an AUC of 87.80%. However, the focus of the study only on classification of DBT into benign or malignant limits its potential for more important multi-class classification, and the relatively low accuracy highlights the need for further improvement.

A SIFT-DBT model was introduced by Du et al. (2024) [16], this model combines self-supervised contrastive learning and patch-level multiple instances learning to address class imbalance. Robust pairs from different slices were formed to focus on structural and semantic information within the images. Their system classified images into normal and abnormal and achieved an AUC of 92.69%, specificity of 84.15%, and sensitivity of 84.62%. However, their focus on binary classification only limits their work. Additionally, while the AUC is high, the specificity and sensitivity suggest that there is still room for performance improvement.

For benign versus malignant DBT classification, Mendes et al. (2023) [17] developed a model that was built using a primary framework based on previous research by Muduli et al. (2021) [18]. The model integrated the well-established data augmentation techniques with the developed CNN architecture that was optimized using the Adam technique for the classification. Their model achieved an accuracy of 93.2%, and an F1-score of 94%. However, the emphasis of the model on binary classification (benign versus malignant) limits its ability for a multi-class classification model that includes normal cases.

A graph convolutional neural network, to classify DBT scans into cancerous and non-cancerous, was proposed by Bai et al. (2022) [19]. They utilized two datasets in this study, one of which is a private and the other is a public dataset. The model merges spatial-based self-attention pooling graph convolution network and graph representation (GCN). Their performance was then compared to that of baseline models such as 3D ResNet,

ResNet-Vote, Two-stream, and Spatial ResNet. The model achieved an accuracy, sensitivity, and an F1-score of 84%, 84%, and 83%, respectively. However, the study is limited by a relatively low accuracy in binary classification.

A study was conducted by Nogay, Akinci, and Yilmaz (2021) [20] for a binary and quadruple classification of DBT scans. The study employed transfer learning techniques on five traditional pre-trained deep convolutional neural network (DCNN) models: ResNet-18, AlexNet, GoogleNet, and ShuffleNet. New weights were assigned to the newly developed layers in the five pre-trained DCNN models, while keeping weights of existing layers unchanged. Accuracy rates ranged from 65 to 75% for binary classification, and from 66 to 86% for quadruple classification.

In this paper, a fully automated system based on deep learning and a classification ensemble was studied. The aim is to develop a multi-class DBT classification system that would be able to efficiently classify the DBT scans into normal, benign, and malignant. The performance of the developed system was compared with the performance of the state-of-the-art deep learning models. In this study, the BCS-DBT [21] public dataset was utilized, and the results were compared with the outcomes of prior research utilizing the same dataset.

2 Methodology

In this study, we introduced MSAE-DL, an integrated comprehensive multi-head self-attention ensemble deep learning system that combines feature fusion, selection, a multi-head self-attention model, and a novel classification ensemble model. As illustrated in Fig. 1, the system processes DBT slices by first applying several augmentation techniques, followed by image enhancement and colour mapping. One of the biggest challenges in deep learning for medical imaging is ensuring the model does not overfit to a limited dataset. To address this, data augmentation techniques were applied to artificially expand the training dataset, helping the model learn more robust and generalizable features. This included random rotations, horizontal flipping, and contrast enhancements, simulating real-world variations that can occur in DBT scans. Additionally, the images were enhanced using histogram equalization and colour mapping techniques, improving contrast and highlighting subtle abnormalities that might otherwise be missed. Subsequently, features are extracted using Mod_AlexNet [22], a previously

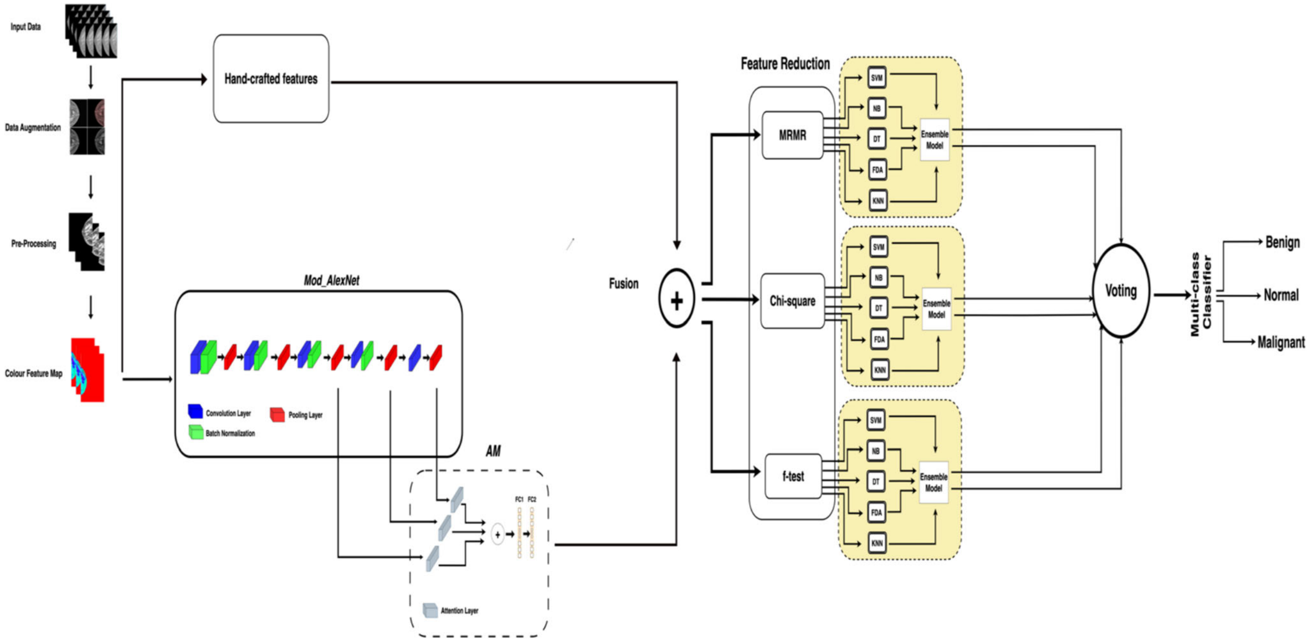


Fig. 1 Diagram of the system developed in the work reported herein

developed deep learning model that achieved outstanding results, consistently surpassing the performance of other state-of-the-art models [22]. By integrating this high-performing model with our newly developed multi-head attention model, we aim to significantly enhance classification, pushing the limits of what current deep learning systems can achieve in DBT scans analysis. Moreover, HOG descriptors are also extracted from the enhanced images. These feature vectors are then combined through concatenation. Following feature fusion, the fused features undergo selection through three high-performing feature selection models to reduce the feature vector by selecting the optimal features from each model. Finally, the selected features are fed into a novel classification ensemble model that incorporates class and classifier weights in a maximum voting ensemble approach to generate the final prediction. The developed multi-class classification system classifies DBT scans into normal, benign, and malignant classes. In addition to data augmentation, other techniques deployed to address overfitting, in the work reported in this paper, include: regularization, feature selection, ensemble methods, and cross-validation.

The proposed system was evaluated on the BCS-DBT dataset [21] with the following performance metrics: accuracy, sensitivity, precision, specificity, and f1-score. The key contributions of the work reported herein may be outlined as follows:

- Develop a novel multi-class classification DBT classification system.
- Evaluate the impact of a multi-head self-attention model in the extraction of optimal features for better discrimination between classes.
- Utilize high-performing feature selection models for feature reduction and removal of redundant features that challenge the classification performance of DBT scans.
- Develop a novel classification ensemble model, to reduce overfitting and improve classification performance compared to single classifiers.
- Evaluate the efficiency of the developed system utilizing a publicly accessible dataset and compare it with prior research utilizing the BCS-DBT dataset and state-of-the-art deep learning models.

2.1 Multi-head self-attention model

The self-attention model is an attention model that weighs and connects different positions of input images before making predictions. The primary goal of the attention model is to highlight the most relevant information while ignoring redundant information about the input image [23]. Multi-head self-attention layers plays a vital role in the image classification tasks, when integrated with deep learning models, due to their capacity to extract complex links and contextual information from feature maps generated by the pooling layers. Multi-head self-attention layers are able to extract discriminative features, thus resulting in enhanced classification performance. The architecture of a multi-head attention module is provided in Fig. 2.

Several learnt parameters and mathematical operations operate the multi-head attention layers. After receiving feature maps from the pooling layers, the attention layers use linear transformations to calculate query, key, and value matrices—typically denoted as Q, K, and V, respectively [23]. The Q, K, and V are the foundational matrices for the evaluation of the relative importance of features and are used to compute the attention scores between pairs of features. The attention score α_{ij} between features i and j is calculated as follows, in Eq. 1:

$$\alpha_{ij} = \mathrm{sof}t\max\left(\frac{Q_i.K_j^T}{\sqrt{d_k}}\right) \tag{1}$$

where Q_i and K_j represent the query and key vectors derived from features i and j, respectively, and d_k is the dimensionality of the key vectors [23]. The architecture of the developed attention model is presented in Fig. 3.

Fig. 2 Architecture of a multi-head attention module [23]

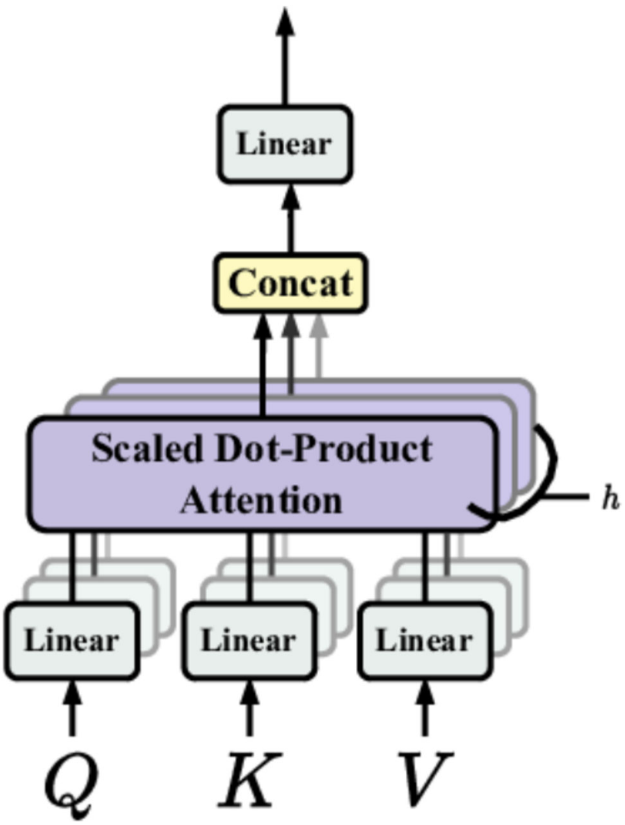
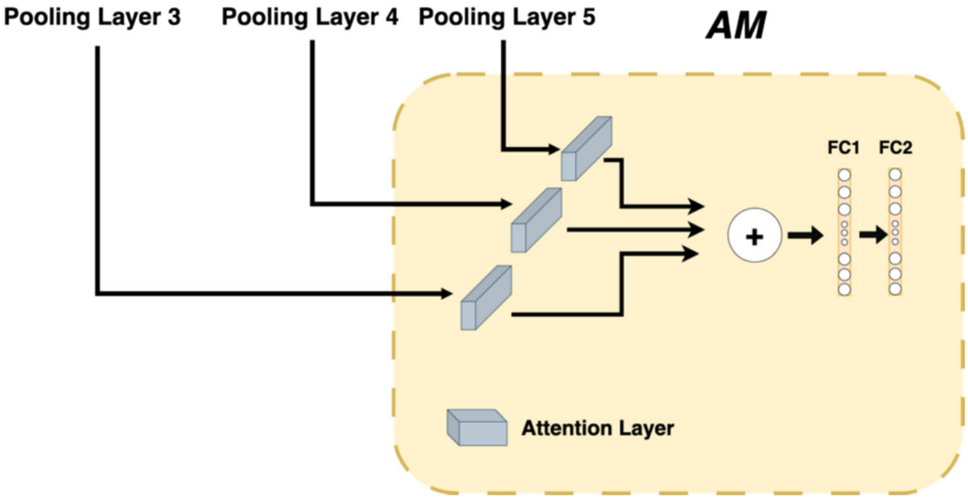


Fig. 3 Architecture of the developed attention model



This developed attention model was integrated with the previously developed Mod_AlexNet model [22]. Mod_AlexNet is an enhanced version of AlexNet developed in our prior research [22]. This model proved to outperform traditional AlexNet and other deep learning models when employed on the BCS-DBT dataset, especially in the classification of abnormal classes [22]. This model adds 6 layers to the original layers of AlexNet, including 2 max-pooling layers, and 4 batch normalization layers. The max-pooling layers were added to the first two convolutional layers, to enhance the extraction of more reliable and relevant features, especially when focusing on low-level features in the earlier layers. This addition helps in identifying low-contrast features and complex anatomical structures available in the DBT scans with varying sizes and locations. On the other hand, the batch normalization layers were added to the first four convolution layers to reduce the internal covariant shift and improve the data flow between the intermediate layers. These layers enhance the ability of the model to handle variations in intensity levels that are a challenge caused by varying tissue composition and X-ray penetration.

The multi-head attention layers were connected to the 3rd, 4th, and 5th pooling layers of the Mod_AlexNet and consist of 8 heads, 64 key and query channels, 256 value channels, and 256 output channels. To be able to extract diverse and more discriminative features, 8 heads were included in this layer. To maintain the ability to capture complex patterns while effectively computing the attention scores, 64 keys and query channels were assigned in this layer. Standardizing parameter selection across self-attention layers reinforces flawless integration and compatibility within network architecture. This results in enhanced computational efficiency, as well as feature extraction process. Afterwards, the outputs of the three self-attention layers are concatenated to produce a combined feature representation. This attention model was trained utilizing several optimizers and on different batch sizes. The optimal performance was demonstrated on a batch size of 64 utilizing the SGDM optimizer.

To improve the stability of the model and ensure it generalizes well to new data, several regularization techniques were incorporated to prevent overfitting. One of the key methods is the batch normalization layer, which helps keep activations stable across different layers and mini-batches. This technique speeds up training by reducing fluctuations in the learning process, making it easier for the model to converge efficiently while maintaining accuracy. To further optimize training different learning strategies, including stochastic gradient descent with momentum (SGDM), Adam, and RMSProp were tested. These optimizers adjust learning rates dynamically, preventing the model from diverging during training. After evaluating their effectiveness, SGDM was chosen as the final optimizer because it provided the best balance between stability and classification accuracy. These techniques, previously introduced in our Mod_AlexNet development [22], are crucial in ensuring that improvements in classification performance are attributed to the robustness of the model rather than over-fitting to the training dataset. This approach strengthens the ability of the model to provide reliable and accurate classifications for DBT scans.

The HOG descriptors break images down into small regions and calculate the direction of the pixel gradients within each region. These gradient directions are then aggregated into histograms, which effectively capture local edge orientations and patterns. In this study, HOG descriptors were extracted to capture fine-grained edge and shape features in DBT scans. Features were extracted from the trained attention model and fused with the HOG descriptors using the concatenation method to be input to the feature selection model. By combining precise edge detection capabilities of HOG with the high-level abstract features extracted from the attention model, we achieved a more comprehensive feature representation. This fusion significantly enhances the robustness and accuracy of the classification task.

2.2 Feature selection

Following feature concatenation, high-performing feature selection techniques were deployed to reduce the risk of overfitting in complex deep learning models, by minimizing the number of features input to the classifiers. Feature selection techniques reduce the dimensionality of the feature vector and enhance the performance of the classifier by retaining informative and relevant features and removing redundant and less informative features. The feature selection stage ensures that classification improvements are due to the effectiveness of the algorithm rather than an overfitted feature space. The three feature reduction strategies used were: minimum redundancy maximum relevance (mRMR), Chi-squared, and f-tests.

Minimum redundancy maximum relevance (mRMR) is a feature selection technique that aims to carefully select a subset of features that integrate high relevance features into the target class with low inter-feature redundancy. Relevance scores are given to each feature based on its relationship with the target class [24]. Max-relevance searches for features satisfying Eq. 2 [24].

$$\max D(S,c), D = \frac{1}{|S|}\sum_{x_i \in S} I(x_i;c) \tag{2}$$

where S is the feature set, c is the sample class, and x_i is the individual feature in the feature set S. The feature is approximated using the mean value $\max D(S,c)$ of mutual information between individual features x_i and class c [24]. Features selected based on max-relevance are likely to have a high level of redundancy, implying a high degree of dependency among these features. When two features are strongly dependent on one another, removing one of them has little impact on the discriminative ability of the other class. Minimum redundancy, shown in Eq. 3, is deployed to remove the redundancy present between two features [24].

$$\min R(S), R = \frac{1}{|S|^2}\sum_{x_i, x_i \in S} I\left(x_i, x_j\right) \tag{3}$$

Features are ranked by concurrently minimizing the redundancy and maximizing relevance as shown in Eq. 4.

$$\max \Phi(D,R), \Phi = \mathrm{D} - \mathrm{R} \tag{4}$$

Chi-squared is a feature selection model that is usually applied to categorical data. This technique determines the degree of independence between each feature and the target class. The Chi-square statistics, also written as χ^2, is calculated, as shown in Eq. 5, using the contingency table. The basic idea behind the Chi-square test, is to evaluate whether the observed distribution of values of a feature significantly differs from the predicted distribution under the null hypothesis of independence between the feature and target variable [25].

$$\chi^2 = \sum \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \tag{5}$$

where O_{ij} indicates the observed frequency of feature value i in class j, and E_{ij} represents the predicted frequency of feature i value in class j, as determined by the independence assumption. The summation is performed over all feature values and class labels. The number of feature values and class labels determines the degree of freedom in the Chi-square statistics.

Finally, the ANOVA (Analysis of Variance) F-test, is a statistical approach to determine the most relevant features for the classification task by comparing several independent means. Features are ranked by calculating the ratio of variances within and between groups [26]. The F-statistic, given in Eq. 6, is calculated as the ratio of the variance between class means to the variance within each class.

$$F = \frac{MSB}{MSW} \tag{6}$$

MSB represents the mean square between classes and computed as the variance of feature values across classes weighted by the number of samples in each class. On the other hand, MSW represents the average variance of the feature values within each class.

2.3 Ensemble model

In this study, a novel ensemble model was developed, as shown in Fig. 4, to improve the classification of abnormal classes. The main contribution was the integration of class and classifier weights for making this model. The stacking ensemble approach is utilized in this model, to combine the predictions of several base classifiers

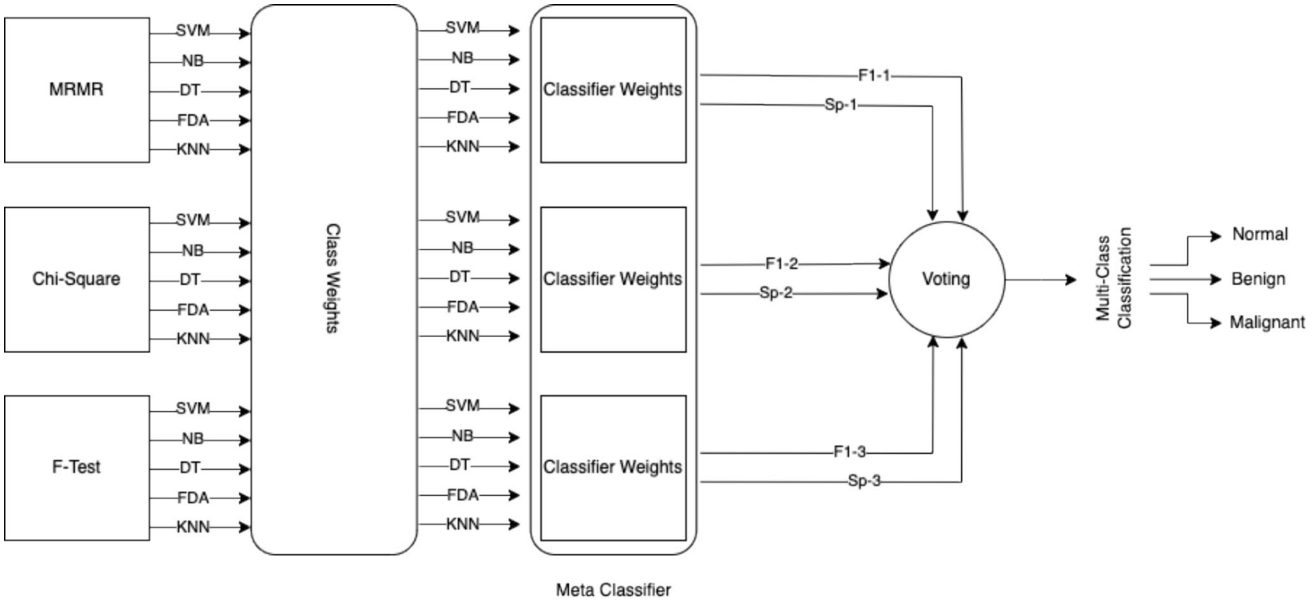


Fig. 4 Architecture of the developed ensemble model

while considering class weights to overcome the weaknesses of individual classifiers and improve the overall predictions. The stacking ensemble model deployed a hierarchical architecture that integrated several classifiers to generate predictions. Each classifier was trained on the same data subset. This diversity is important due to the complex nature of abnormalities found in tomosynthesis scans. Each classifier produces predictions and is input to the meta-classifier which combines them and produces the final prediction.

The integrated novel ensemble model aims to enhance the predictions by adding more accurate weights for classifiers and classes. Specifically, the class weights address the class imbalance challenge that is faced using the considered dataset. Given the unbalanced distribution of normal, benign, and malignant cases in the dataset, adjusting the class weights assists in overcoming the imbalance by giving more weight to underrepresented classes. This strategic change strengthens the discriminative ability of the classifier and guarantees an equal impact for each class in the decision-making process. Class weights are calculated using the formula presented in Eq. 7.

$$w_i = \frac{n_s}{n_c * n_{si}} \tag{7}$$

w_i represents the weight assigned to each class, with i indicating the specific class. The numerator n_s represents the total number of samples in the dataset, to provide a measure of the overall size of the dataset. On the other hand, n_c denotes the total number of unique classes within the target variable, illustrating the diversity of classes present in the dataset. Finally, n_{si} provides the total number of instances associated with the respective class i. This formula adjusts the weights according to the frequency of each class compared to the whole dataset size. The integration of class weights aims to enhance the sensitivity of the classification model to the individual features of each class, by assigning higher weights to classes that have less instances and lower weights to classes with more instances, resulting in a more balanced training and learning process.

Following the assignment of class weights for the prediction of each classifier, the predictions are input to the meta-classifier that represents an advancement in the classification process. By fusing classifier predictions with performance metrics such as specificity and the F1-score, this novel model builds a sophisticated framework for generating weighted outcomes. The F1-score is a complete assessment of classifier performance that combines precision and recall, whereas specificity assesses the ability to properly recognize true negatives. Considering these performance metrics into the integrated ensemble model improves the performance of the classification model. This meta-classifier presents a new decision-making framework through the integration of performance indicators and classifier predictions. In this study, several weighted predictions for each class from various classifiers, including SVM, NB, DT, FDA, and KNN, using multiple feature selection techniques, utilize the collective intelligence embedded in various prediction techniques. Finally, a maximum vote ensemble model is implemented to combine the predictions generated by each base classifier and feature selection model, selecting the class label with the most votes.

Figure 4 presents the developed integrated framework, where three feature selection models produce classifier predictions, which are then weighted by a class weight model. Predictions are then refined by a meta-classifier by assigning weights based on the f1-score and specificity of the classifier. The final prediction is achieved through a maximum voting framework. This introduced framework offers advantages over traditional ensemble methods by incorporating performance measures into the weighting system, leading to better decision-making. It provides precise control over the weighting mechanism, resulting in more accurate and consistent predictions. The following section describes the dataset utilized, performance measures, system implementation, results, and discussion; giving a thorough overview of its deployment and outcomes.

3 Materials and methods

3.1 Dataset

Selecting the right dataset is crucial for deep learning in medical applications, especially when dealing with complex models like CNNs. In this study, we used the Breast Cancer Screening-Digital Breast Tomosynthesis (BCS-DBT) [21] public dataset which contains three-dimensional scans. The dataset was collected by Duke University Hospital/Duke University in Durham, North Carolina, USA. The BCS-DBT dataset includes a total of 22,032 DBT scans collected from 5,060 patients. During a typical Digital Breast Tomosynthesis (DBT) examination, patients generally have two scans per breast, one from the top (craniocaudal or CC view) and one from the side (mediolateral oblique or MLO view), totalling four scans. Since each scan can produce between 40 and 100 images, a patient might end up with anywhere from 160 to 400 images from the entire examination. The dataset was made up using the evaluations that were performed between August 2014 and January 2018. The dataset is classified into four categories: actionable (non-biopsied)—further imaging examination was recommended, normal, benign, and malignant based on biopsy results. Only the normal, biopsy-proven benign, and biopsy-proven malignant categories were taken into consideration for the purposes of this study. The BCS-DBT dataset is the only publicly available resource specifically tailored for research on Digital Breast Tomosynthesis (DBT) scans, making it an essential foundation for progress in this field. While obtaining large-scale annotated medical datasets is always a challenge, BCS-DBT offers a diverse and clinically relevant set of cases, making it an excellent choice for developing and evaluating breast cancer detection systems. Consequently, it has become the primary dataset used in both existing and new studies for the classification and detection of abnormalities, ensuring consistency and reliability in DBT research.

Table 1 provides a breakdown of the distribution of cases among the categories in the BCS-DBT dataset for the training, testing, and validation sets. To be able to handle the large number of cases in the dataset, dataset size,

Table 1 Statistics of the BCS-DBT dataset [21]

Number of Patients	Training	Validation	Testing	Total
Normal	4109	200	300	4609
Actionable	178	40	60	278
Benign	62	20	30	112
Malignant	39	20	30	89

and bias towards the normal cases, only cases for 600 patients were considered in this study. When increasing the number of patients cases in the study, the class imbalance in the dataset increases and impacts the performance of the deep learning models. This occurs because real-world datasets typically mirror the natural distribution of conditions, with normal cases being far more common than benign or malignant cases. Consequently, adding new cases, primarily from the majority class, disproportionately enhances its representation, leading to an imbalance in the dataset. When training deep learning models, they may exhibit bias towards the majority class, in this case, normal cases, resulting in a limited capacity to predict the minority classes—in this case, benign and malignant cases. The disparity arises from the unequal distribution within the initial dataset, where 96% comprises normal cases and only 4% includes both benign and malignant cases.

In this study, the number of patients in each class is as follows: 499 belong to the normal class, 62 to the benign class, and 39 to the malignant class. When classifying DBT patients, the actionable group—group of individuals who received follow-up imaging due to the identification of a mass or deformity in the study report, but chose not to proceed with a biopsy, were excluded from the study. Due to the necessity of further imaging for evaluating their status, establishing a dependable classification for this particular subgroup is unachievable.

To ensure the integrity and validity of our system, the entire case (all related scans) for each patient was assigned to a single group—either training, validation, or testing. This approach prevents any overlap of scans from the same patient across multiple groups, thereby preventing potential bias that could arise from correlated images. To ensure that the system learns general patterns rather than patient-specific features, we keep scans from the same patient in one group. In this study, we allocated 80% of the data for model training and cross-validation, with this portion undergoing tenfold cross-validation. The remaining 20% of the data was set aside as a distinct test set, completely excluded from both the training and cross-validation phases. The developed system was set up using MATLAB R2023a. The implementation was performed using a 2.8 GHz Quad-Core, Intel Core i7, 16 GB RAM, 1 TB Storage, and an Intel Iris Plus Graphics 655 (1536 MB).

3.2 Performance measures

The developed system classified DBT scans into normal, benign, and malignant. To examine the performance of the developed system on the BCS-DBT dataset, several performance measures were considered, namely, accuracy, sensitivity, precision, specificity, and F1-score. Each measure was computed individually for every class, and the final value was calculated using the weighted average technique. Eqs 1 – 5 provide each measure for a single class.

While classifying, a TP (True Positive) refers to correct positive predictions, where real positive cases were classified as positive. TN (True Negative) refers to correct negative predictions, where real negative cases were classified as negative. FP (False Positive) refers to incorrect positive predictions, where real negative cases are classified as positive. Finally, FN (False Negative) refers to incorrect negative predictions, where real positive cases are classified as negative.

Accuracy, shown in Eq. 8, is the portion of the total samples that were correctly classified by the classifier.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

Sensitivity, shown in Eq. 9, refers to the percentage of real positives that are correctly classified.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{9}$$

Precision, shown in Eq. 10, calculates the proportion of correct positive predictions among all predictions identified as positive.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{10}$$

Specificity, shown in Eq. 11, measures the proportion of real negatives that are correctly classified.

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{11}$$

Finally, the F1-score, shown in Eq. 12, is the combined average of precision and sensitivity to provide a balanced metric.

$$\text{F1 - Score} = \frac{2{*}\text{Precision}{*}\text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} \tag{12}$$

4 Results

The developed system was deployed to enhance the multi-class classification of DBT scans. The performance of the developed system was evaluated and compared, with a specific highlight on accurately classifying the abnormal class, which constitutes the minority class in the utilized dataset. The developed system integrates a novel self-attention model with feature fusion and selection models, alongside a novel ensemble model, to enhance classification performance. System performance was also compared to that of previous work that utilized the same dataset.

Following the data augmentation process, the images were enhanced during the pre-processing stage. Histogram equalization was utilized to adjust pixel intensities, thereby enhancing contrast and improving the visibility of subtle features in the DBT images. To further refine the image quality and suppress unwanted artefacts, Gaussian smoothing was applied as a noise reduction technique. Figure 5 provides examples of normal, benign, and malignant cases, after these pre-processing methods were implemented.

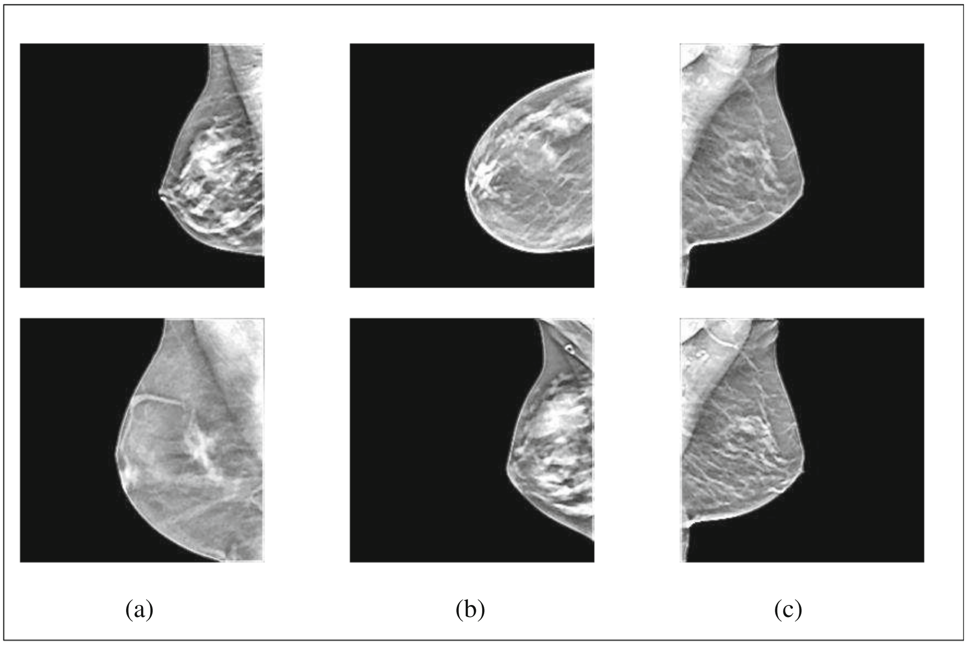


Fig. 5 Samples of **a** Benign cases **b** Malignant cases **c** Normal cases, after pre-processing

To support the differentiation of tissue types within the scan, the HSV colour feature map was integrated after pre-processing. This step leverages the unique properties of the HSV colour space, hue for dominant colours, saturation for intensity, and brightness for lightness, ensuring that critical visual details are captured. The enhanced colour information provided by the HSV feature map enriches the input data for the classification model, improving its capacity to accurately identify and distinguish between normal, benign, and malignant breast tissues. Figure 6 demonstrates the effectiveness of this approach by showcasing examples from each tissue type post application of the feature map.

The initial stage of the system was the development and training of the introduced self-attention model. The model was incorporated with Mod_AlexNet and trained utilizing the SGDM optimizer at a learning rate of 0.0001, 50 epochs, and a batch size of 64. The training and validation curves for loss and accuracy were plotted for 50 epochs and are shown in Figs. 5 and 6.

Figure 7 illustrates the training and validation accuracy curves across 50 epochs. Starting with a modest training accuracy of 45%, the curve increases rapidly, eventually stabilizing at a peak of 99%. Notably, the training accuracy achieves a near-constant level of accuracy from the 20th epoch onward. On the other hand, the validation accuracy commences at 65% and steadily climbs, culminating in a peak accuracy of 93%. Following the 25th epoch, the validation accuracy also approaches a nearly steady state.

In Fig. 8, presenting the training and validation loss, the training loss initiated at 2.46 and steadily declined, reaching a stable 0.004 by the 20th epoch, maintaining nearly this loss until the 50th epoch. Conversely, the validation loss started at 2.16 and progressively decreased, achieving its minimum loss of 0.01 by the 20th epoch, remaining relatively consistent thereafter until the 50th epoch.

Following the feature extraction from the trained attention model, the features extracted were fused with the HOG descriptors. This fusion of features aimed to improve the generalization of the data, employing both the learnt attention-based features and conventional HOG descriptors. Table 2 analyses the outcomes on several performance measures from the attention model and the fusion stage.

As depicted in Table 2, classifying features extracted solely from the attention model demonstrated outstanding performance across all metrics, achieving 90.99%, 91.98%, 51.99%, and 91.26% in terms of accuracy, precision, specificity, and F1-score, respectively. When considering fusing these features with HOG descriptors and classifying them, the fused features achieve an accuracy of 90.89%, a precision of 93.55%, a specificity of 53.55%, and an F1-score of 91.40%. Precision, specificity, and f1-score show improvement when fusing the attention

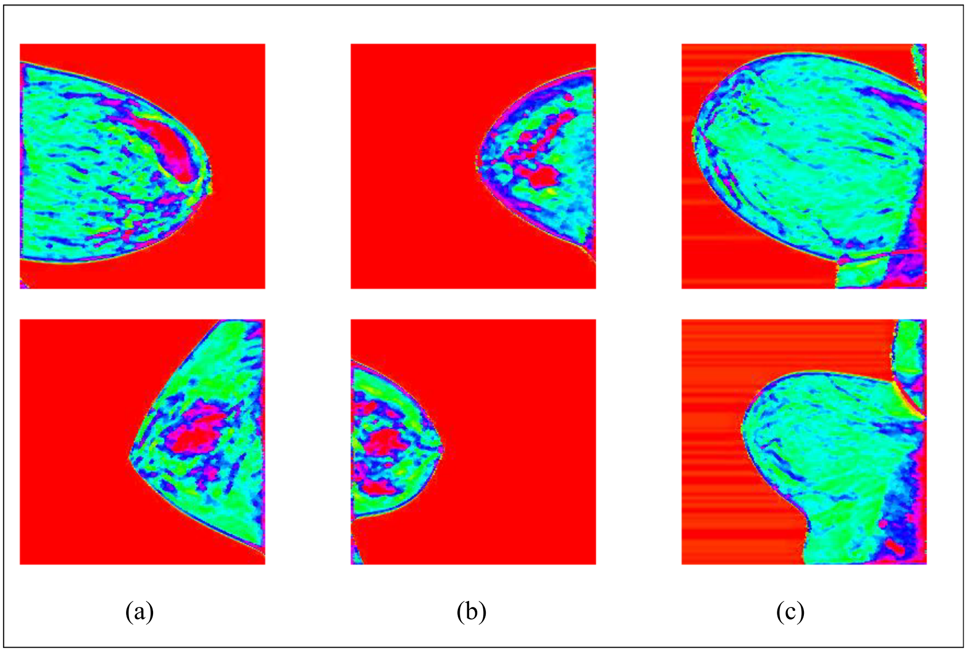


Fig. 6 Samples of **a** Benign cases **b** Malignant cases **c** Normal cases, after the colour mapping technique

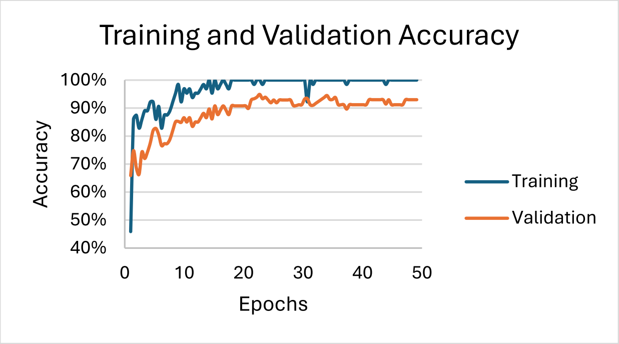


Fig. 7 Self-attention training and validation accuracy versus the number of epochs

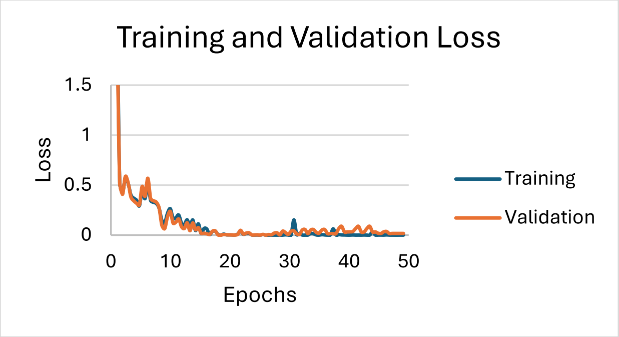


Fig. 8 Self-attention training and validation loss versus the number of epochs

Table 2 Performance assessment of the implementation of attention and fusion models

Model	Performance Measure				
	Accuracy %	Sensitivity/Recall %	Precision %	Specificity %	F1-score %
Self-Attention	90.99	90.99	91.98	51.99	91.26
Fusion	90.89	90.89	93.55	53.55	91.40

features with HOG descriptors, contributing to improved performance and classification capability, particularly in the abnormal classes.

The output of the feature fusion was further processed through three different feature selection methods: mRMR, Chi-square test, and f-test. Each selected set of features was evaluated using several classifiers: NB, SVM, DT, FDA, and KNN. The performance of each feature selection method with each classifier is presented in Tables 3, 4, and 5. These results are compared with those obtained from features extracted by integrating a self-attention model with Mod_AlexNet.

Table 3 presents a comprehensive evaluation of several classifiers utilizing the mRMR feature selection model in comparison to the baseline integrated self-attention Mod_AlexNet model. Using the mRMR feature selection model, the accuracy ranges from 89.16% to 93.10%, with KNN exhibiting the highest accuracy at 93.10%, closely

Table 3 Performance evaluation of employing the mRMR feature selection model with various classifiers

Different integrated contexts	Classifier	Performance Measure				
		Accuracy	Sensitivity	Precision	Specificity	F1-score
SelfAttention_ModAlexNet	90.99%	90.99%	91.98%	51.99%	91.26%	–
Feature Selection (MRMR)	NB	89.16%	89.16%	91.03%	50.36%	89.94%
	SVM	90.41%	90.41%	93.73%	55.07%	91.16%
	DT	89.23%	89.23%	92.31%	57.15%	90.37%
	FDA	92.99%	92.99%	92.82%	45.17%	92.54%
	KNN	93.10%	93.10%	93.81%	53.88%	92.97%

Table 4 Performance evaluation of employing the Chi-square test feature selection model with various classifiers

Different integrated contexts	Classifier	Performance Measure				
		Accuracy	Sensitivity	Precision	Specificity	F1-score
SelfAttention_ModAlexNet	90.99%	90.99%	91.98%	51.99%	91.26%	–
Feature Selection (Chi-square test)	NB	91.12%	91.12%	93.49%	53.19%	91.57%
	SVM	90.44%	90.44%	93.66%	55.43%	91.18%
	DT	89.23%	89.23%	92.31%	57.15%	90.37%
	FDA	92.77%	92.77%	92.51%	47.09%	92.41%
	KNN	91.39%	91.39%	93.41%	53.47%	91.77%

Table 5 Performance evaluation of employing the f-test feature selection model with various classifiers

Different integrated contexts	Classifier	Performance Measure				
		Accuracy	Sensitivity	Precision	Specificity	F1-score
SelfAttention_ModAlexNet	90.99%	90.99%	91.98%	51.99%	91.26%	–
Feature Selection (f-test)	NB	88.62%	88.62%	90.87%	50.19%	89.56%
	SVM	90.38%	90.38%	93.61%	54.40%	91.12%
	DT	88.39%	88.39%	92.87%	57.98%	90.02%
	FDA	92.54%	92.54%	92.66%	48.10%	92.34%
	KNN	93.55%	93.55%	93.87%	53.74%	93.31%

followed by FDA at 92.99%. On the other hand, the precision, which assesses the accuracy of positive predictions, ranges from 91.03% to 93.81%, with KNN again achieving the highest precision at 93.81%, closely followed by SVM at 93.73%. The specificity, an important measure that represents the proportion of actual negatives correctly identified, ranged from 45.17% to 57.15%, with FDA achieving the lowest specificity. The FDA demonstrated exceptional performance in accuracy, sensitivity, precision, and f1-score. With an accuracy of 92.99% and a precision of 92.82%, FDA exhibits proficiency in accurately classifying both positive and negative cases. On the other hand, the FDA achieves the lowest specificity of 45.17%, suggesting a higher rate of false positives. The SVM performs competitively across most metrics, particularly excelling in precision with a value of 93.73%. However, its accuracy at 90.41%, is slightly lower compared to KNN and FDA. The specificity measured for SVM is 55.07%; it outperforms the specificity achieved by the baseline model. DT achieves an accuracy of 89.23%, which is lower than the baseline model; it also achieves the highest specificity at 57.15%, outperforming all classifiers and the baseline model. The higher specificity of DT indicates its ability to better detect negative cases.

Table 4 provides an assessment of various classifiers utilizing the Chi-square test feature selection model and compares their performance to the baseline integrated self-attention Mod_AlexNet model. FDA achieved the highest accuracy at 92.77%, followed by KNN at 91.39% and NB at 91.12%. Moreover, in terms of precision, SVM achieves the highest precision at 93.66%, followed by NB at 93.49% and KNN at 93.41%. This signifies the proficiency of SVM in accurately classifying normal cases. Decision Tree (DT) emerges with the highest specificity at 57.15%, outperforming other classifiers and the baseline model. This suggests the ability of DT to better detect abnormal classes compared to other classifiers. Moreover, SVM achieves a 55.43% specificity, second highest, while FDA demonstrates the least specificity at 47.09%.

Table 5 provides an analysis of the performance of various classifiers when employing the f-test feature selection model. When classifying using the NB classifier, accuracy of 88.62% was achieved, which is somewhat lower than that of SVM. Moreover, NB obtained a low specificity value among all classifiers (50.19%). The SVM, on the other hand, exhibited a higher accuracy of 90.38% compared to NB. SVM recorded a precision of 93.61% and a specificity of 54.40%, which is the second highest precision and sensitivity among other classifiers and outperforms the baseline model. Meanwhile, DT achieves an accuracy of 88.39%, comparable to NB and lower than that achieved by the baseline model. However, it achieves a precision of 92.87% which is slightly higher than the baseline model. Moreover, DT records the highest specificity of 57.98% which outperforms the other classifiers and the specificity of the baseline model. Although FDA outperformed the baseline model, NB, SVM, and DT in terms of accuracy, precision, and F1-score by achieving 92.54%, 92.66%, and 92.34%, respectively, it achieved the lowest specificity of 48.10%. Finally, KNN emerges as the top performer with the highest accuracy of 93.55%, indicating a significant improvement over the baseline and other classifiers. KNN also demonstrates strong performance across other metrics, including sensitivity, precision, and f1-Score. KNN achieved a specificity of 53.74% which is not the highest among other classifiers, but higher than the baseline model.

After applying the feature selection models, the newly developed ensemble model is utilized. By intelligently merging the strengths of various classifiers while addressing their individual limitations, we developed a powerful voting stacking ensemble model. This ensemble model aims to achieve better predictions than any single classifier by leveraging the collective knowledge of several classification models. The output from the meta-classifier includes two predictions generated by each feature selection model: one based on the F1-score for each classifier and the other based on the specificity measure. To arrive at the final prediction, the output of the meta-classifier is processed through a maximum voting model, as shown in Table 6.

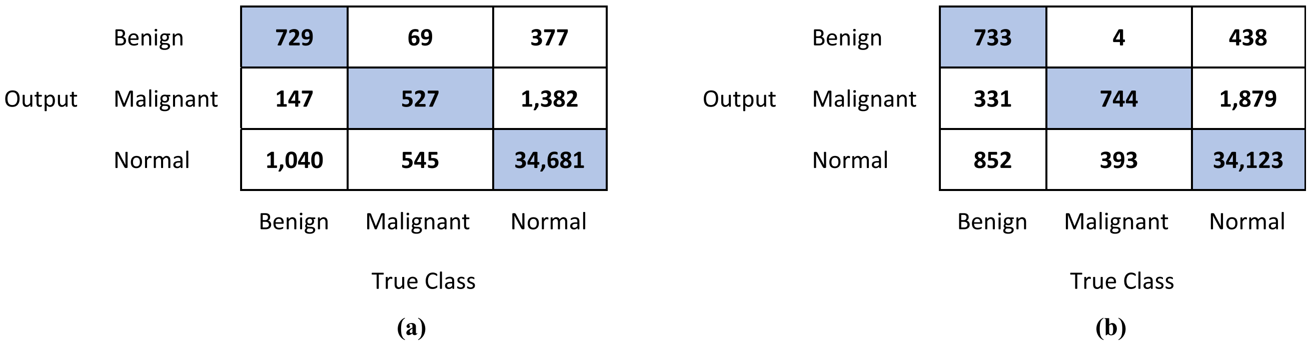
When analysing the results from Table 6, the ensemble model utilizing the mRMR feature selection shows notable improvement, particularly in the F1-score, reaching 92.02%. This superior performance is driven by its selection of features that maximize relevance while minimizing redundancy, thus enhancing predictive power. Similarly, the Chi-square ensemble model achieved an F1-score of 91.58% by selecting features based on their statistical significance, though its specificity of 53.10% suggests further room for improvement. The final ensemble model demonstrated an accuracy of 90.13% and an improved specificity of 62.20%, attributable to its integration of the multiple classifiers, which effectively reduced the misclassification rates.

Figure 9 shows the confusion matrices for the SelfAttention_ModAlexNet system (a) and the MSAE-DL system (b), comparing their classification performance. The SelfAttention_ModAlexNet system correctly identified 729 benign cases, 527 malignant cases, and 34,681 normal cases. In comparison, the MSAE-DL system slightly improved the detection of benign cases, increasing the number of correct classifications to 733, and it significantly enhanced the detection of malignant cases, raising true positives from 527 to 744. However, this improvement came with a slight drop in accuracy for normal cases, where correctly classified instances decreased from 34,681 to 34,123. These results highlight the ability of the MSAE-DL system to better detect malignant cases while balancing its performance across other classes**.**

The comparison between the SelfAttention_ModAlexNet and the final ensemble model MSAE-DL is presented in Table 7. Initially, SelfAttention_ModAlexNet exhibits a slightly higher accuracy of 90.99% compared to the MSAE-DL system accuracy of 90.13%. However, the MSAE-DL system outperforms the SelfAttention_ModAlexNet, achieving a precision of 92.77% compared to the precision of the SelfAttention_ModAlexNet

Table 6 Performance evaluation of the developed ensemble model

Different integrated contexts	Performance Measure					
	Accuracy	Sensitivity/Recall	Precision	Specificity	F1-score	–
SelfAttention_ModAlexNet	90.99%	90.99%	91.98%	51.99%	91.26%	–
Ensemble for MRMR	F1	91.76%	91.76%	93.31%	53.42%	92.02%
	Specificity	91.30%	91.30%	93.15%	53.97%	91.72%
Ensemble for Chi-square	F1	91.12%	91.12%	93.35%	53.10%	91.58%
	Specificity	91.07%	91.07%	93.39%	53.67%	91.57%
Ensemble for f-test	F1	91.69%	91.69%	93.37%	53.42%	91.99%
	Specificity	91.23%	91.23%	93.39%	53.52%	91.69%
Final Ensemble	90.13%	90.13%	92.77%	62.20%	91.03%	



(a)

		True Class		
		Benign	Malignant	Normal
Output	Benign	**729**	**69**	**377**
	Malignant	**147**	**527**	**1,382**
	Normal	**1,040**	**545**	**34,681**

(b)

		True Class		
		Benign	Malignant	Normal
Output	Benign	**733**	**4**	**438**
	Malignant	**331**	**744**	**1,879**
	Normal	**852**	**393**	**34,123**

Fig. 9 **a** Confusion matrix for the SelfAttention_ModAlexNet system; **b** Confusion matrix for the MSAE-DL system

Table 7 Comparison of performance between SelfAttention_ModAlexNet and MSAE-DL

Different integrated contexts	Performance Measure				
	Accuracy %	Sensitivity/Recall %	Precision %	Specificity %	F1-score %
SelfAttention_ModAlexNet	90.99	90.99	91.98	51.99	91.26
MSAE-DL	90.13	90.13	92.77	62.20	91.03

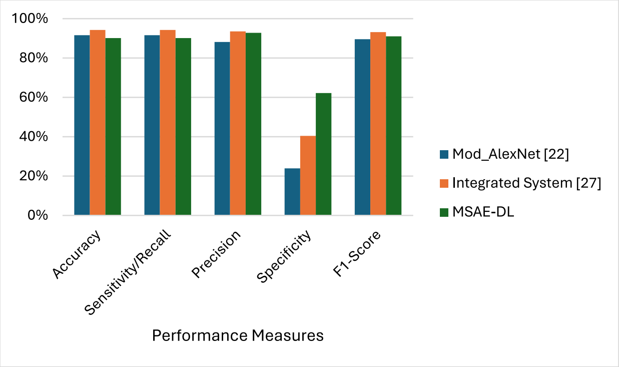
which is 91.98%. This indicates that the MSAE-DL system is more effective at correctly classifying true positives among those predicted as positive. The higher precision of the MSAE-DL system is attributed to its ability to incorporate different feature selection models and the developed ensemble model that capture more discriminative features, relevant to positive cases. Furthermore, the specificity measure demonstrates a considerable advantage for the MSAE-DL system over the SelfAttention_ModAlexNet, with a specificity of 62.20% versus 51.99%. This shows that the MSAE-DL system excels at correctly recognizing true negatives among the occurrences predicted as negative, demonstrating better discrimination between positive and negative cases. The improved specificity of the MSAE-DL system is attributed to its ability to efficiently classify abnormal classes. Furthermore, the SelfAttention_ModAlexNet achieved a slightly higher F1-score of 91.26% compared to 91.03% achieved by MSAE-DL. In summary, while the SelfAttention_ModAlexNet achieved a slightly higher accuracy, the MSAE-DL system has a significant advantage in precision and specificity. The increased specificity of MSAE-DL indicates an improved ability to accurately classify abnormal classes. This demonstrates that the integration of the developed SelfAttention_ModAlexNet with the selected feature selection models and the newly developed ensemble model in the effectiveness of the MSAE-DL system in establishing the complex structure of feature space, resulting in better discrimination between normal and abnormal classes.

Throughout our research, we utilized a consistent dataset for both training and testing phases to facilitate accurate comparisons. To assess the performance of the MSAE-DL system relative to our previously developed system [22, 27], we ensured that the training and testing sets were identical. Table 8 and Fig. 10 provide a comparative analysis of the MSAE-DL system alongside the systems developed earlier. Mod_AlexNet [22] achieved an accuracy of 91.61% and demonstrated a low specificity of 23.91% that indicates that a substantial

Table 8 Comparison of our MSAE-DL system with our previous work

Our work	Performance Measure				
	Accuracy %	Sensitivity/Recall %	Precision %	Specificity %	F1-score %
Mod_AlexNet [22]	91.61	91.61	88.16	23.91	89.57
Integrated System [27]	94.27	94.27	93.51	40.42	93.13
MSAE-DL	90.13	90.13	92.77	62.20	91.03

Fig. 10 Comparison of our MSAE-DL system with our previous work



number of abnormal cases are incorrectly classified as normal. The Integrated System [27], which incorporates Mod_AlexNet for feature extraction fused with HOG descriptors and followed by mRMR feature selection, yielded superior results with an accuracy of 94.27% and a specificity of 40.42%. The incorporation of the HOG descriptors enhances the ability of the system to capture critical edge and shape information, improving the discrimination. In comparison, the MSAE-DL system achieved an accuracy of 90.13%, which, while slightly lower than that of the Integrated System, is accompanied by a notable specificity of 62.20%. This suggests that the use of a multi-head self-attention model, integrated with the novel ensemble model, enhanced the feature extraction and classification. The self-attention mechanism enabled the model to weigh different features dynamically, emphasizing those that contribute more significantly to the classification model. Additionally, the ensemble model integrated predictions from multiple classifiers, which further reduces the risk of misclassifications of abnormal cases. The results highlight that the MSAE-DL system demonstrates a notable improvement in specificity, achieving 62.20% compared to 23.91% for the Mod_AlexNet system [22] and 40.42% for the Integrated System [27]. This indicates that MSAE-DL is more effective at correctly identifying negative cases, a critical aspect in reducing false positives. This suggests that the MSAE-DL system improved in correctly identifying "Benign" and "Malignant" cases, leading to enhanced specificity. However, this advancement comes with a slight increase in misclassification of "Normal" cases as either "Benign" or "Malignant", likely because the model focuses more on distinguishing subtle features in the minority classes.

Many studies have examined the classification of DBT scans using the BCS-DBT dataset, often integrating it with some private dataset, which can lead to variability in results. Classification strategies vary including distinctions between benign and malignant cases, normal versus abnormal (where "abnormal" includes both benign and malignant), and cancerous versus non-cancerous (where "non-cancerous" encompasses benign and normal cases). To ensure fair comparisons, multiple iterations of our MFSAE-DL system were conducted by modifying the classification approach. Our system was tested in three previously mentioned configurations, and the results are presented and compared with previous work in Tables 9, 10, and 11.

Table 9 shows that our system performed significantly better than the model by Nogay, Akinci, and Yilmaz (2021) [20], achieving 90.13% accuracy compared to their 75.00% in the multi-class classification. This is largely due to the more advanced design of our model, which uses a multi-head self-attention mechanism and an ensemble classification approach. These features improved how well the system could extract important details from the data and combine the strengths of different classifiers. In contrast, the model developed by Nogay et al.

Table 9 Results of multi-class classification for normal, benign, and malignant classes

Author/Year	Accuracy	Sensitivity/Recall	Precision	Specificity	F1-score
Our System (MSAE-DL)	90.13%	90.13%	92.77%	62.20%	91.03%
Nogay, Akinci, and Yilmaz, (2021) [20]					

Table 10 Results of classification for cancerous versus non-cancerous classes

Author/Year	Accuracy %	Sensitivity/Recall %	Precision %	Specificity %	F1-score %	AUC
Our System (MSAE-DL)	93.81	94.01	99.60	87.20	96.72	0.91
Tardy and Mateus (2021) [28]						0.73
Nogay, Akinci, and Yilmaz, (2021) [20]	86.00					
Bai et al. (2022) [19]	84.00	84.00	86.00		83.00	
Adhikesaven et al. (2022) [29]	97.25					
Bai et al. (2022) [30]	92.00	93.00	91.00	91.00	92.00	

Table 11 Results of classification for normal versus abnormal

Author/Year	Accuracy %	Sensitivity/Recall %	Precision %	Specificity %	F1-score	AUC
Our System (MSAE-DL)	94.53	95.83	98.20%	79.03	97.00%	0.87
Du et al. (2024) [16]		84.62		84.15		0.92
Fogleman, Otsap, and Cho (2021) [31]	94.90					

relied on pre-trained networks with transfer learning, which may not have been as effective in handling the complexities of DBT images, resulting in lower accuracy in multi-class classification.

In comparing the results for the classification of cancerous versus non-cancerous cases, in Table 10, various systems demonstrate different strengths. Our system achieved high performance with an accuracy of 93.81%, a sensitivity of 94.01%, a precision of 99.60%, a specificity of 87.20%, an F1-score of 96.72%, and an AUC of 0.91, due to the use of a multi-head self-attention mechanism and ensemble classification. Tardy and Mateus (2021) [28] reported an AUC of 0.73, but their model likely faced challenges due to the complexity of DBT images and reliance on a private multi-vendor dataset. Nogay et al. (2021) [20] achieved 86.00% accuracy using pre-trained DCNNs, but their model lacked the advanced feature extraction methods of our system. Bai et al. (2022) [19] reported an accuracy of 84.00% using graph convolutional networks, and their feature fusion Siamese network [30] achieved 92.00% accuracy, 93.00% sensitivity, and 91.00% precision, due to its innovative comparison of current and prior mammograms. Finally, Adhikesaven et al. (2022) [29] achieved the highest accuracy at 97.25% with a CNN for early detection, though the lack of precision and specificity metrics makes direct comparison challenging.

When comparing the results from Table 11, for the classification of normal versus abnormal cases, our system demonstrated a high accuracy of 94.53%, with sensitivity at 95.83%, precision of 98.20%, specificity of 79.03%, an F1-score of 97.00%, and an AUC of 0.87. In contrast, Du et al. (2024) [16] reported an accuracy of 84.62% and a higher AUC of 0.92. Despite their use of a novel self-supervised initialization and fine-tuning method (SIFT-DBT) for imbalanced data, the lower accuracy could be attributed to the difficulty of managing data imbalance through their patch-level multi-instance learning approach. Similarly, Fogleman, Otsap, and Cho (2021) [31] achieved a slightly higher accuracy of 94.90% with their system, which utilized transfer learning with a partial Inception v3 architecture. However, their lack of reported metrics for specificity and precision limits the depth of direct comparison with our system.

Only a few researchers have exclusively utilized the BCS-DBT subset only without the integration of a private dataset. When comparing our work with studies that used only the BCS-DBT dataset, we utilized the same data and adapted our MFSAE-DL system to match their classification models, ensuring fairness in our comparisons and validations. The results are presented in Tables 12 and 13.

Table 12 Comparison of classification results for benign versus malignant cases using the BCS-DBT dataset only (Scenario 1)

Author/Year	Accuracy %	Sensitivity/Recall %	Precision %	Specificity	F1-score %
Our System (MSAE-DL)	91.09	92.74	88.49	89.73%	90.45
Hassan et al. (2022) [32]	80.43				
Hassan et al. (2024) [33]	85.00	90.00	84.10		86.90

Table 13 Comparison of classification results for benign versus malignant cases using the BCS-DBT dataset only (Scenario 2)

Author/Year	Accuracy%	Sensitivity/Recall %	Precision	Specificity %	F1-score
Our System (MSAE-DL)	91.24	91.35	89.69%	91.14	90.40%
Farangis Sajadi Moghadam and Rashidi (2023) [34]	88.67	77.12		75.11	
Farangis Sajadi Moghadam and Rashidi (2024) [15]	78.51	82.78		75.19	

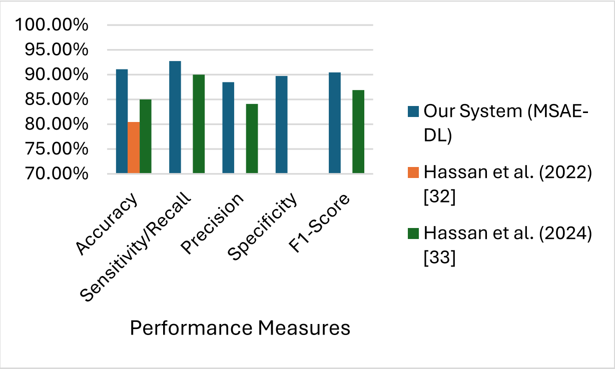


Fig. 11 Performance comparison for benign versus malignant case classification using the BCS-DBT dataset (Scenario 1)

When comparing the results from Table 12, as shown in Fig. 11, our system achieved an accuracy of 91.09%, with a sensitivity of 92.74%, precision of 88.49%, specificity of 89.73%, and an F1-score of 90.45%. These metrics demonstrate that our model effectively classified benign and malignant cases, maintaining a strong balance between sensitivity and specificity. In contrast, Hassan et al. (2022) [32] reported a lower accuracy of 80.43%, likely due to limitations in their deep learning-based radiomics approach combined with SVM classification. The performance of their model was possibly hindered by a smaller training dataset and the use of traditional machine learning techniques like SVM, which may have restricted feature extraction. In their later work, Hassan et al. (2024) [33] achieved a higher accuracy of 85.00%, along with a sensitivity of 90.00%, precision of 84.10%, and an F1-score of 86.90%. This improvement was likely due to the introduction of image quality-aware features and tumour texture descriptors, which enhanced the feature extraction capabilities of the model. However, it still fell short compared to our system. Overall, our system outperformed both versions of models developed by Hassan et al.

When comparing the results in Table 13, as shown in Fig. 12, our system demonstrated superior performance, with an accuracy of 91.24%, sensitivity of 91.35%, precision of 89.69%, specificity of 91.14%, and an F1-score of 90.40%. These metrics highlight the effectiveness of our model, which leverages advanced feature extraction and classification techniques to accurately distinguish between benign and malignant cases. In contrast, Farangis Sajadi Moghadam and Rashidi (2023) [34] achieved a lower accuracy of 88.67%, with sensitivity at 77.12% and specificity at 75.11%. Their approach, which employed radiomic-based feature extraction and Quadratic Discriminant Analysis (QDA), was effective but fell short in terms of sensitivity and precision, likely due to

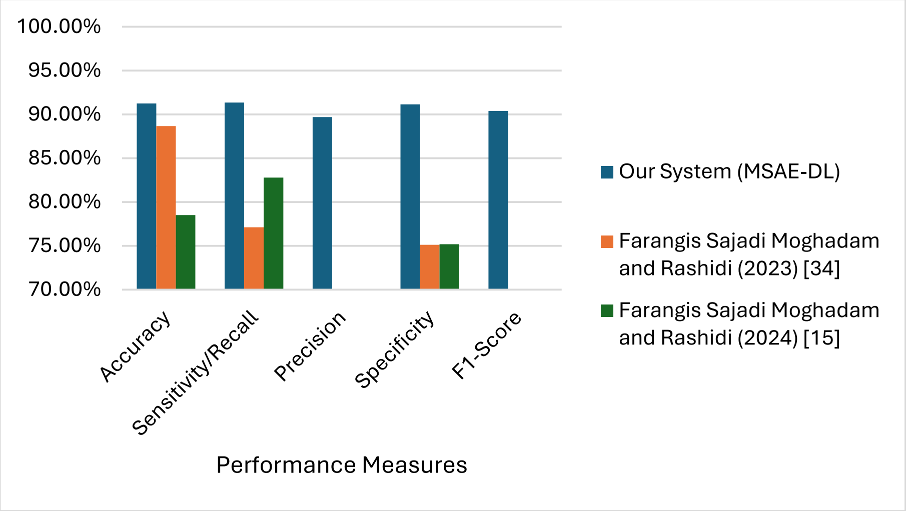


Fig. 12 Performance comparison for benign versus malignant case classification using the BCS-DBT dataset (Scenario 2)

limitations in feature extraction and model selection, which may not have fully captured the complexity of DBT images.

In their subsequent study (2024) [15], Farangis Sajadi Moghadam and Rashidi reported an even lower accuracy of 78.51%, with sensitivity of 82.78% and specificity of 75.19%. Although they introduced a novel feature extraction method based on DCT-DOST features, the performance of the model remained constrained. The relatively lower precision and specificity suggest that the system may have overfitted the training data, particularly given the smaller sample sizes, which likely hindered its ability to generalize to new data. Overall, our system outperformed both of their studies, especially in terms of accuracy, sensitivity, and specificity.

Despite the lower performance in specificity achieved by our developed MSAE-DL system, it is important to understand this statistic within the framework of multi-class classification. Unlike the majority of prior research, which focused mainly on binary classification scenarios, the MSAE-DL system is designed for multi-class classification, as shown in Tables 8 and 9. The only comparable study is that of Nogay, Akinci, and Yilmaz [20], as they also classified images into benign, malignant, and normal classes using the BCS-DBT dataset. However, our system demonstrates superior performance compared to theirs, demonstrating its efficacy in this multi-class classification task.

5 Discussion

DBT is a three-dimensional imaging technique which effectively reduces false positives and negatives caused by overlapping breast tissue in traditional 2D mammography. Research data from clinical trials suggests that computer-aided detection (CAD) systems have the potential to improve breast cancer detection and assist radiologists in their diagnostic evaluations, thus improving overall diagnostic accuracy. In accordance with these findings, a multi-diagnostic system is proposed, which integrates a customized self-attention deep learning architecture with feature selection and a novel ensemble model.

In this study, images from the BCS-DBT dataset were augmented and enhanced followed by application of two feature extraction models. Features were extracted using the SelfAttention_ModAlexNet developed in our work, alongside HOG descriptors. These features were then fused using a concatenation technique. Subsequently, three feature selection models were employed to identify the most relevant and informative features for enhancing classifier performance. The resultant feature sets from the feature selection models were input into the novel ensemble model. Finally, images were classified into three categories: normal, benign, and malignant.

Experimental results indicate that while individual models may enhance performance, integrating all models and constructing the entire system leads to superior performance compared to traditional CNN models and the baseline model. In this study, a novel multi-head self-attention model was developed to enhance the performance of the previously developed Mod_AlexNet [22], incorporating three multi-head self-attention layers to enhance the feature extraction phase. Furthermore, an ensemble model was developed, integrating class and classifier weights, and applying them to predict the outcome for each classifier using three deployed feature selection models.

Various performance indicators were measured when comparing the performance of the MSAE-DL system developed in our work. This system achieved an accuracy of 90.13%, a precision of 92.77%, a specificity of 62.20%, and an f1-score of 91.03%, surpassing our previously developed systems [22] [27] and other traditional deep learning models and prior research in the multi-class classification of DBT scans.

Although the MSAE-DL system achieved high overall accuracy and precision, the relatively lower specificity (62.20%) indicates that there is still room for improvement, particularly in classifying abnormal cases. Moreover, another aspect to consider is slice selection during the processing of DBT scans. Selecting the most informative slices is crucial to enhance classification performance of the system.

Thus, while this study introduces a highly effective multi-class classification system that outperforms existing models, there are areas to address in future work. These include improving specificity and optimizing slice selection.

6 Conclusion and future work

The focus on the classification of Digital Breast Tomosynthesis (DBT) by machines has been limited, in the available literature to date, exposing a considerable need for comprehensive research. One of the main challenges to progress in this field is a lack of publicly available datasets. Previous research has mostly focused on developing binary classification systems by classifying the DBT scans into normal or abnormal, ignoring the need for a more complicated classification model. Furthermore, most studies in this field have failed to investigate the complexities of a multi-class classification system to classify the scans into normal, benign, and malignant. This traditional and commonly used binary classification approach is considered overly simplistic, ignoring the potential benefits of a more complex three-class classification approach. Implementing a three-class classification system increases the accuracy and specificity of DBT scan diagnosis, providing radiologists with a useful medical tool for better patient care.

Our study introduces an innovative multi-class classification system incorporating a self-attention deep learning model, feature fusion and selection techniques, and a novel ensemble classification model. This system effectively optimizes and combines features extracted from our developed SelfAttention_ModAlexNet model with HOG descriptors. Subsequently, the fused features are input to three feature selection models, producing the most relevant and informative feature sets. These sets are then fed into a developed ensemble model, which integrates class and classifier weights, assigning them to predictions and classifiers for each feature selection model.

The results demonstrate that our proposed system, MSAE-DL, outperforms both our previously developed systems [22, 27] and prior research into multi-class classification systems using the BCS-DBT dataset. Various performance metrics, including accuracy, sensitivity, precision, specificity, and f1-score, were assessed. Our MSAE-DL system achieved an accuracy of 90.13%, a precision of 92.77%, a specificity of 62.20%, and an f1-score of 91.03%, surpassing the performance of Nogay, Akinci, and Yilmaz [20], who achieved an accuracy of 75%. Notably, their study represents the only prior research utilizing the considered dataset to develop a multi-class DBT classification system.

Although our work, reported in this paper, has made significant progress in multi-class classification of DBT scans, it still faces some limitations. It is specifically designed for DBT imaging, and it does not incorporate multimodal radiological data, which may offer additional diagnostic insights. In future work, we envision several enhancements of our proposed MSAE-DL system. Firstly, we plan to enhance the feature selection model to extract more relevant and discriminative features. Furthermore, we aim to investigate the potential integration of other deep learning architectures and techniques into our system to further boost the performance. Finally, we are interested in expanding and augmenting the BCS-DBT dataset or acquiring additional datasets to enhance the robustness of the training data of our system. This proposed future work will be explored to provide clinicians with actionable insights and facilitate the integration of our system into real-world clinical workflows.

Acknowledgements No acknowledgments were made for this research.

Author contributions A.M.A.E.-S., S.M.Y., A.H.S., and C.C. contributed to conceptualization, methodology, resources, validation, and visualization. A.M.A.E.-S helped in software, formal analysis, investigation, and writing—original draft preparation. S.M.Y. and A.H.S. helped in writing—review and editing. S.M.Y., A.H.S., and C.C. helped in supervision and project administration. All authors have read and agreed to the published version of the manuscript.

Funding Open access funding provided by The Science, Technology & Innovation Funding Authority (STDF) in cooperation with The Egyptian Knowledge Bank (EKB). This research received no external funding.

Data availability The dataset used in this research is publicly available through The Cancer Imaging Archive. The data used is the “Breast Cancer Screening–Digital Breast Tomosynthesis (BCS-DBT) (Version 5)” dataset (Buda et al., 2020). It can be accessed at https://doi.org/10.7937/E4WT-CD02. This dataset is not owned by the authors and is used under the terms and conditions specified by the original data providers.

Declarations

Conflict of interest The authors declare that they have no competing interests.

References

1. Ferlay J, Ervik M, Lam F, Colombet M, Mery L, Piñeros M, Znaor A, Soerjomataram I, Bray F (2021) Global Cancer Observatory: Cancer Today. Lyon, France: International Agency for Research on Cancer. https://gco.iarc.fr/today
2. Mittra I, Mishra GA, Dikshit RP, Gupta S, Kulkarni VY, Shaikh HK, Shastri SS, Hawaldar R, Gupta S, Pramesh CS, Badwe RA (2021) Effect of screening by clinical breast examination on breast cancer incidence and mortality after 20 years: prospective, cluster randomised controlled trial in Mumbai. BMJ. https://doi.org/10.1136/bmj.n256
3. Helvie MA (2010) Digital mammography imaging: breast tomosynthesis and advanced applications. Radiologic Clinics 48:917–929. https://doi.org/10.1016/j.rcl.2010.06.009
4. Liu W, Wang Z, Liu X, Zeng N, Liu Y, Alsaadi FE (2017) A survey of deep neural network architectures and their applications. Neurocomputing 234:11–26. https://doi.org/10.1016/j.neucom.2016.12.038
5. Gao M, Samala RK, Fessler JA, Chan HP (2020) Deep convolutional neural network denoising for digital breast tomosynthesis reconstruction. Medical Imaging 2020: Physics of Medical Imaging. SPIE 11312:173–178. https://doi.org/10.1117/12.2549361
6. Gao M, Fessler JA, Chan HP (2021) Deep convolutional neural network with adversarial training for denoising digital breast tomosynthesis images. IEEE Trans Med Imaging 40:1805–1816. https://doi.org/10.1109/tmi.2021.3066896
7. Su T, Deng X, Yang J, Wang Z, Fang S, Zheng H, Liang D, Ge Y (2021) DIR-DBTnet: Deep iterative reconstruction network for three-dimensional digital breast tomosynthesis imaging. Med Phys 48:2289–2300. https://doi.org/10.1002/mp.14779
8. Sulaiman SN, Normazli MH, Harron NA, Karim NK, Ahmad KA, Soh ZH (2022) A Convolutional Neural Network Model for Image Enhancement of Extremely Dense Breast Tissue in Digital Breast Tomosynthesis Images. IEEE 12th

International Conference on Control System, Computing and Engineering (ICCSCE). https://doi.org/10.1109/iccsce54767.2022.9935647

9. Saifudin SA, Sulaiman SN, Karim NK, Osman MK, Isa IS, Harron NA (2022) A comparative study of unsharp masking filters for enhancement of digital breast tomosynthesis images. IEEE 12th International Conference on Control System, Computing and Engineering (ICCSCE). https://doi.org/10.1109/iccsce54767.2022.9935638
10. Mota AM, Mendes J, Matela N (2023) Digital Breast Tomosynthesis: Towards Dose Reduction through Image Quality Improvement. Journal of Imaging 9:119. https://doi.org/10.3390/jimaging9060119
11. Visvkarma AK, Sehra K, Laishram R, Malik A, Sharma S, Kumar S, Rawal DS, Vinayak S, Saxena M (2022) Impact of gamma radiations on static, pulsed I-V, and RF performance parameters of AlGaN/GaN HEMT. IEEE Trans Electron Devices 69:2299–2306. https://doi.org/10.1109/ted.2022.3161402
12. Chen X, Lv J, Wang Z, Qin G, Zhou Z (2024) Deep-AutoMO: Deep automated multiobjective neural network for trustworthy lesion malignancy diagnosis in the early stage via digital breast tomosynthesis. Comput Biol Med 183:109299
13. Shao Z, Cai Y, Hao Y, Hu C, Yu Z, Shen Y, Lu H (2024) AI-based strategies in breast mass$\leq$ 2 cm classification with mammography and tomosynthesis. The Breast 78:103805
14. Oladimeji OO, Ayaz H, McLoughlin I, Unnikrishnan S (2024) Mutual information-based radiomic feature selection with SHAP explainability for breast cancer diagnosis. Results in Engineering 24:103071. https://doi.org/10.1016/j.rineng.2024.103071
15. Farangis Sajadi Moghadam and Rashidi S (2024) Novel feature extraction based on DCTDOS features for classification of Digital Breast Tomosynthesis images into benign and malignant tumors. Research Square. https://doi.org/10.21203/rs.3.rs-3931625/v1
16. Du Y, Hooley RJ, Lewin J, Dvornek NC (2024) SIFT-DBT: Self-supervised Initialization and Fine-Tuning for Imbalanced Digital Breast Tomosynthesis Image Classification. arXiv. https://doi.org/10.48550/arxiv.2403.13148
17. Mendes J, Matela N, Garcia N (2023) Avoiding Tissue Overlap in 2D Images: Single-Slice DBT Classification Using Convolutional Neural Networks. Tomography 9:398–412. https://doi.org/10.3390/tomography9010032
18. Muduli D, Dash R, Majhi B (2022) 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
19. Bai J, Jin A, Jin A, Wang T, Yang C, Nabavi S (2022) Applying graph convolution neural network in digital breast tomosynthesis for cancer classification. In Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics 1–10. https://doi.org/10.1145/3535508.3545549
20. Nogay HS, Akinci TC, Yilmaz M (2021) Comparative experimental investigation and application of five classic pre-trained deep convolutional neural networks via transfer learning for diagnosis of breast cancer. Advances in Science and Technology Research Journal 15:1–8. https://doi.org/10.12913/22998624/137964
21. Buda M, Saha A, Walsh R, Ghate S, Li N, Święcicki A, Lo JY, Mazurowski MA (2020) Detection of masses and architectural distortions in digital breast tomosynthesis: a publicly available dataset of 5,060 patients and a deep learning model. arXiv. https://arxiv.org/abs/2011.07995
22. El-Shazli AM, Youssef SM, Soliman AH (2022) Intelligent Computer-aided model for efficient diagnosis of digital breast tomosynthesis 3D imaging using Deep Learning. Appl Sci 12:5736. https://doi.org/10.3390/app12115736
23. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Advances in neural information processing systems 20. https://doi.org/10.1109/TPAMI.2005.159
24. Peng H, Long F, Ding C (2005) Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27:1226–1238. https://doi.org/10.1109/TPAMI.2005.159
25. McHugh ML (2013) The chi-square test of independence. Biochemia medica 23:143–149
26. Nasiri H, Alavi SA (2022) A novel framework based on deep learning and ANOVA feature selection method for diagnosis of COVID-19 cases from chest X-ray images. Comput Intell Neurosci. https://doi.org/10.1155/2022/4694567
27. El-Shazli AM, Youssef SM, Soliman AH, Chibelushi C (2024) An Enhanced Framework Employing Feature Fusion for Effective Classification of Digital Breast Tomosynthesis Scans. In 2024 International Conference on Machine Intelligence and Smart Innovation (ICMISI). IEEE
28. Tardy M, Mateus D (2021) Trainable summarization to improve breast tomosynthesis classification. International Conference on Medical Image Computing and Computer-Assisted Intervention 140–149
29. Adhikesaven S, Kapoor A, Khowaja AS, Li V, Sabhanayakam K, McMahan L (2022) Predicting the Instance of Breast Cancer within Patients using a Convolutional Neural Network. Journal of Emerging Investigators. https://doi.org/10.59720/22-061
30. Bai J, Jin A, Wang T, Yang C, Nabavi S (2022) Feature fusion Siamese network for breast cancer detection comparing current and prior mammograms. Med Phys 49:3654–3669. https://doi.org/10.1002/mp.15598
31. Fogleman S, Otsap J, Cho S (2021) Clinical Diagnosis Support with Convolutional Neural Network by Transfer Learning. SMU Data Science Review 5

32. Hassan L, Abdel-Nasser M, Saleh A, Puig D (2022) Breast Tumor Classification in Digital Tomosynthesis Based on Deep Learning Radiomics. In Artificial Intelligence Research and Development 269–278
33. Hassan L, Abdel-Nasser M, Saleh A, Puig D (2024) Classifying Breast Tumors in Digital Tomosynthesis by Combining Image Quality-Aware Features and Tumor Texture Descriptors. Machine Learning and Knowledge Extraction 6:619–641
34. Moghadam FS, Rashidi S (2023) Classification of benign and malignant tumors in Digital Breast Tomosynthesis images using Radiomic-based methods. In 2023 13th International Conference on Computer and Knowledge Engineering (ICCKE) 203–208. IEEE

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

Authors and Affiliations

Alaa M. Adel El-Shazli[1] · **Sherin M. Youssef**[1] · **Abdel Hamid Soliman**[2] · **Claude Chibelushi**[3]

✉ Alaa M. Adel El-Shazli
alaa.alshazli@aast.edu; alaa.elshazli@research.staffs.ac.uk

Sherin M. Youssef
sherin@aast.edu

Abdel Hamid Soliman
a.soliman@staffs.ac.uk

Claude Chibelushi
claude.chibelushi@semantics21.com

[1] Arab Academy for Science, Technology and Maritime Transport (AASTMT), Giza, Alexandria, Egypt

[2] Staffordshire University, Stoke-on-Trent, UK

[3] Semantics Ltd, Southampton, UK