Computers in Biology and Medicine 169 (2024) 107914

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Fusion of transfer learning models with LSTM for detection of breast cancer using ultrasound images

Madhusudan G. Lanjewar [a,*], Kamini G. Panchbhai [b], Lalchand B. Patle [c,**]

[a] School of Physical and Applied Sciences, Goa University, Taleigao Plateau, Goa, 403206, India
[b] Goa College of Pharmacy, Panaji, Goa, 403001, India
[c] PG Department of Electronics, MGSM's DDSGP College Chopda, KBCNMU, Jalgaon, Maharashtra, 425107, India

ARTICLE INFO

Keywords:
Breast cancer
MobileNetV2
VGG16
Long short-term memory LSTM
SMOTETomek
Ultrasound images
Gradient-weighted class activation mapping (Grad-CAM)

ABSTRACT

Breast Cancer (BC) is one of the top reasons for fatality in women worldwide. As a result, timely identification is critical for successful therapy and excellent survival rates. Transfer Learning (TL) approaches have recently shown promise in aiding in the early recognition of BC. In this work, three TL models, MobileNetV2, ResNet50, and VGG16, were combined with LSTM to extract the features from Ultrasound Images (USIs). Furthermore, the Synthetic Minority Over-sampling Technique (SMOTE) with Tomek (SMOTETomek) was employed to balance the extracted features. The proposed method with VGG16 achieved an F1 score of 99.0 %, Matthews Correlation Coefficient (MCC) and Kappa Coefficient of 98.9 % with an Area Under Curve (AUC) of 1.0. The K-fold method was applied for cross-validation and achieved an average F1 score of 96 %.

Moreover, the Gradient-weighted Class Activation Mapping (Grad-CAM) method was applied for visualization, and the Local Interpretable Model-agnostic Explanations (LIME) method was applied for interpretability. The Normal Approximation Interval (NAI) and bootstrapping methods were used to calculate Confidence Intervals (CIs). The proposed method achieved a Lower CI (LCI), Upper CI (UCI), and Mean CI (MCI) of 96.50 %, 99.75 %, and 98.13 %, respectively, with the NAI, while 95 % LCI of 93.81 %, an UCI of 96.00 %, and a bootstrap mean of 94.90 % with the bootstrap method. Furthermore, the performance of the six state-of-the-art (SOTA) TL models, such as Xception, NASNetMobile, InceptionResNetV2, MobileNetV2, ResNet50, and VGG16, were compared with the proposed method.

1. Introduction

Recent studies show that BC is the most prevalent malignancy among women worldwide, with 2.3 millions new cases recognized yearly [1,2], and 10 % of women have BC [3]. The statistics on BC in India published in 2022 highlight that 15 % of the deaths are caused by BC and growing at a rate of 13 % every five years [4]. Developing cancer cells in breast tissue is a complicated process that includes several variables. Although the precise aetiology of BC is not yet entirely known, some risk factors have been found. These risk factors include age, family history of BC, specific genetic mutations, hormonal variables, and way of life [5]. BC has also been linked to reproductive variables, such as early menstrual start, delayed menopause, never having children, or having them later in life [6,7]. Moreover, the environmental variables linked to the development of BC include exposure to radiation and certain chemicals. Advances in health science research also imply that regional variations in risk factors may exist. For instance, research on migration has revealed that women's risk of BC gradually increases as they relocate from low-incidence to high-incidence nations [8].

Early detection of BC is essential for effective treatment and improving the chances of survival. Fortunately, several BC diagnostic methods can aid in the disease's early identification. BC can be identified using Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), Ultrasound (US), and mammography [2]. Mammography is one of the most popular ways to screen for BC [9]. Due to the high quality of the images produced, mammography is believed to be the most effective imaging method for identifying and diagnosing BC. By detecting BC early, mortality from BC can be reduced, and lives can be saved. In the US imaging method, sound waves produce pictures of the breast and provide more details about the features of a worrisome

* Corresponding author.
** Corresponding author.
E-mail addresses: madhusudan@unigoa.ac.in (M.G. Lanjewar), kaaminilanjewar@gmail.com (K.G. Panchbhai), lalchandpatle85@gmail.com (L.B. Patle).

https://doi.org/10.1016/j.compbiomed.2023.107914
Received 30 September 2023; Received in revised form 14 December 2023; Accepted 27 December 2023
Available online 3 January 2024

lesion. Mammography benefits significantly from adding these imaging modalities, particularly in certain diagnostic circumstances. Computer-aided radiologists can use diagnosis tools to help them analyze mammographic images [10]. Increased treatment choices and better survival rates for those with BC are directly related to these improvements in detection methods. In addition to imaging techniques, clinical breast examination is another essential tool for detecting BC [11]. Generally, the most accurate way to diagnose BC uses various imaging methods, including mammography, Ultrasound, MRI, and computer-aided diagnosis tools [12].

The emergence of big data and the growth of data-driven disciplines such as Deep Learning (DL) or Machine Learning (ML) have been used for prediction or classification in various fields [13]. Researchers and healthcare professionals may analyze massive volumes of data and find patterns that might not be visible to the human eye by utilizing the power of ML/DL algorithms. Healthcare experts seek to increase BC diagnosis accuracy and efficiency using ML/DL models, resulting in better patient survival rates. ML/DL models have often been used in healthcare analysis, including BC diagnosis [14]. These algorithms can decipher intricate datasets and find minute patterns that can point to the existence of BC. The author fused TL models with the LSTM and used them as feature extractors. These extracted features were balanced using SMOTETomek and classified using various classifiers. The fusion approach, which combines two or more methods, was used in various domains, including disease detection [15]. The main contributions are:

- ❖ We presented a unique technique integrating TL models, LSTM networks, the SMOTETomek algorithm, and classifiers to construct a reliable and efficient diagnostic tool.
- ❖ As the basis for feature extraction from the US, we use SOTA pre-trained TL models, such as MobileNetV2, ResNet50, and VGG16. This method uses these networks' capacity to acquire hierarchical features important for BC lesions as benign, malignant, or normal.
- ❖ We used LSTM networks to capture US temporal relationships and sequences successfully. Because LSTM units are suitable for extracting information from sequential data, they are suitable for identifying small changes and variances that indicate the development of cancerous tissue.
- ❖ MobileNetV2, ResNet50, and VGG16 with LSTM were employed as feature extractors, automatically extracting features. The informative elements extracted by these TL models were fused with the LSTM approach. The hyper-parameters of these TL models were tuned with an optimizer with their learning rate to obtain the optimal features.
- ❖ We incorporate the SMOTETomek algorithm to counteract the shortage of positive samples and mitigate the effects of class imbalance. SMOTETomek to over-sample the minority class while concurrently under-sampling the majority class. This re-balancing strategy improves the model's capacity to acquire information from both categories, resulting in better generalization and a lower probability of misclassification.
- ❖ To categorize the BC using the TL-LSTM-derived features, several ML classifiers were used. These classifiers' hyper-parameters were fine-tuned, which aids in classification.
- ❖ Combining these strategies, we want to improve the performance of BC diagnosis using US imaging. Not only does our suggested approach improve diagnostic performance, but it also assures resilience over a wide range of patient demographics and imaging situations.
- ❖ Various performance assessment criteria were used to assess the models' efficiency, and the models were cross-validated utilizing the K-fold approach, which contributed to assuring the models' resilience.
- ❖ The Grad-CAM and LIME were applied for visualization, which helped to solve the difficulty of comprehending and believing complicated suggested models, which aids in interpreting individual predictions.
- ❖ The NAI approach was used to obtain population parameters and test hypotheses. It provides a robust framework for statistical inference, allowing researchers to make educated judgments while explaining the uncertainty in predictions. The bootstrapping approach was utilized to generate firm conclusions, build CIs, and solve issues from small sample numbers and non-standard information distributions.
- ❖ To compare the effectiveness of the presented method, six SOTA models, Xception, NASNetMobile, InceptionResNetV2, MobileNetV2, ResNet50, and VGG16, were assessed.

2. Literature survey

Various medical research and bio-medicine fields are exploring ML/DL to detect various diseases. Hajipour Khire Masjidi et al. [16] proposed an ML-based method to identify benign, malignant, and healthy from 1200 US images. They employed a two-dimensional contourlet, and using the time-dependent technique, they featured contourlet sub-bands from the images. The Decision Tree (DT) obtained the highest accuracy of 88.90 % among the other methods for the malignant category. In another study, Liu et al. [17] presented a Deep Neural Network (DNN) based on a grid-based deep feature generator technique to identify BC using ultrasonographic images. The input picture employed was separated into rows and columns for the feature extraction, and these Deep Feature Generators (DFG) were employed in every row and column. After calculating the error value, the best three feature vectors were chosen using the grid-based DFG. They obtained 97.18 % accuracy, an F1 score of 96.79 %, and a Geometric mean of 96.15 % with 10-fold cross-validation. Complex DL models are prone to overfitting the training data, mainly when the dataset is small or needs more variety. A Disease-Specific Imaging (DSI) technique was proposed by Baek et al. [18] for the BC diagnosis. They used a Modified Fully Convolutional Network (MFCN) for the segmentation and a modified GoogleNet to classify breast lesions. Additionally, they carried out a multi-parametric analysis inside the contoured lesions. They used Support Vector Machine (SVM) generated features and B-mode-driven generated features as "post-processing" for DL. The classification accuracy and the AUC were obtained as 98.2 % ± 2.4 % and 94.0 % ± 8.9 %, respectively. Furthermore, VGG19 model was used by Boulenger et al. [19] for the automated recognition of Triple-Negative Breast Cancer (TNBC) purely from USIs. The t-distributed Stochastic Neighbor Embedding and saliency maps were employed for model visualization. They achieved an average AUC of 0.858 (95 % CI: 64 %, 95 %), an accuracy of 85 %, a sensitivity of 86 %, and an F1-score of 74.31 %. Gu Y. et al. [20] used a multi-centre dataset to build a VGG19 with External Test Cohorts (ETC) for discriminating benign from malignant breast tumours from USIs. They achieved an AUC of 91.30 %, sensitivity of 88.84 %, accuracy of 86.40 %, F1 score of 87.20 %, and MCC of 72.8 %. AUC in the comparison set was comparable to the expert's with a p-value of 0.5629. However, the performance could have been better, including the p-value.

Sirjani et al. [21] used InceptionV3 to categorize ultrasonography breast lesions. They used a dataset of 3 types of 2D breast USIs. The model's Root Mean Squared Error (RMSE), recall, F1 score, accuracy, AUC, and Cronbach's alpha values in the test group were 0.18, 77 %, 80 %, 81.0 %, 81.0 %, and 0.77, respectively. However, the performance was poor. In another study, Chen et al. [22] proposed a GoogleNet-based method to categorize benign, malignant, and normal to diagnose BC and achieved an accuracy of 96.37 % and a loss of 0.3492. To forecast female BC, Breast Imaging Reporting and Data System (BI-RADS) US descriptors were proposed by Shen et al. [23]. BD-Net on the test set predicted BC with a specificity of 91.0 %, a sensitivity of 93.8 %, and an AUC of 0.92. However, developing functional features from BI-RADS Ultrasound descriptors necessitates domain knowledge. Liao et al. [24] created an ensemble Entity Discovery and Linking (EDL-BC) to detect BC lesions early. The model was trained on 7955 colour Doppler USIs. The EDL-BC obtained a sensitivity of 94.4 %, and the AUC of the inner and two

separate outward validation cohorts of EDL-BC was 0.950 and 0.907, respectively. Radiologists using AI aid and the EDL-BC had considerably higher AUCs of 0.945 and 0.899 for correct diagnosis than radiologists working alone with an AUC of 0.716. However, the approach's usefulness may be restricted to the lesions seen in the training data. Wang et al. [25] built a DL Network (DLN) with the addition of an Automatic Segmentation Network (ASN) for BC diagnosis using 769 USIs of breast tumours. The ResNet34v2, ResNet50v2, and ResNet101v2 retrieved breast tumour morphological information and appended ASN to the traditional ResNet. The ResNet34v2 obtained a specificity of 76.81 %, accuracy of 78.11 %, Positive Predictive Value (PPV) of 82.22 %, and AUC of 0.85. A DL-based technique was presented by Taleghamar et al. [26] to predict the response of BC to Neoadjuvant Chemotherapy (NAC) employing Quantitative Ultrasound (QUS) multiparametric imaging. The best feature maps from the parametric images were extracted using the Residual Network (RN) and Residual Attention Network (RAN). The RAN achieved an accuracy of 88 % and an AUC of 0.86 on a separate test. Moreover, Fleury and Marcomini [27] classified breast masses using radiomic characteristics from the BI-RADS on US B-mode images. With the help of five ML techniques, they retrieved the ten essential attributes and classified them as malignant or benign. Using 10-fold cross-validation, SVM achieved AUC (0.840), sensitivity (71.4 %) and specificity (76.9 %). Radiologists' interpretations of BI-RADS classifications might differ. This unpredictability can impact the accuracy and repeatability of radiomic feature extraction.

Atrey et al. [28] proposed a semi-automated multimodal breast tumour classification method by fusing characteristics from US and mammography images. They achieved an accuracy of 98.84 % with cubic SVM for the multimodal Computer-Aided Detection (CAD) system. ML combined with quantitative USI characteristics for the detection of Triple-Negative (TN) BC was examined by Wu et al. [29]. They achieved an AUC of 0.88, a sensitivity of 86.96 % and a specificity of 82.91 %. Zhang et al. [30] used a clinical model, radio mics model, and radio mics nomogram with clinical factors to predict BC from 1014 patients with axillary lymph node positive BC. They achieved an AUC of 0.882, an accuracy of 84.6 %, and the external test cohort with an AUC of 0.858 and an accuracy of 77.8 %. Karthiga & Narasimhan [31] used VGG16, AlexNet, ResNet50, VGG19, and ResNet101 to predict BC from USIs. The ten significant features were chosen and obtained an accuracy of 99 % with the K-Nearest Neighbours (KNN) Algorithm, and DCNN obtained an accuracy of 100 % for two class problems. For the three class problems, the accuracy results of VGG16, AlexNet, ResNet50, VGG-19, and Resnet-101 were 82.0 %, 86.0 %, 84.0 %, 84.0 % and 74.0 %, respectively, which could have been better. Romeo et al. [32] proposed an ML-based method to classify non-cystic benign and malignant breast tumours from USIs. The Random Forest (RF) obtained an accuracy of 82 % and an AUC of 0.90. A similar study was proposed by Lu et al. [33] by modifying the AlexNet model with batch normalisation layers to identify aberrant brain activity in MRI. An extreme learning machine then replaced the final layers, and the chaotic bat method was used to improve the classification accuracy of the extreme learning machine. They attained 97.14 % sensitivity, 95.71 % specificity, and 96.43 % total accuracy [33]. Table 1 summarizes the literature-reported work related to BC identification.

Based on the literature described above, a few scholars have begun recognizing BC using various ML and DL methods and have published acceptable outcomes. However, it is still essential to develop practical and highly accurate approaches. We have proposed a hybrid technique for detecting BC to fill the gaps in earlier studies. The fundamental purpose of the suggested research is to determine BC by utilizing a hybrid approach that combines non-destructive image processing with DL and may be applied in real-world circumstances. Imbalanced datasets were utilized to create the system in this work. As a result, the data sampling approach was used to balance the dataset. Rather than using traditional image processing approaches, we applied a DL approach to automatically extract discriminating textural suggestive characteristics. These DL-LSTM-derived features provide a framework for assessing BC in real time.

Table 1
literature-reported work related to BC identification.

References	Dataset	Classes	Methods	Performances	Cross-validation
Hajipour Khire Masjidi et al. [16]	780 BCE Ultrasound (BCUS) images	3	KNN, SVM, DT, RF, and LDA	DT: Accuracy = 88.90 %	Not applied
Liu et al. [17]	780 BCUS images collected from Behaye Hospital	3	Grid-Based DL Model	Accuracy = 97.18 % and F1 score = 96.79 %	10- fold accuracy = 90.38 %
Baek et al. [18]	121 BCUS images collected from University of Rochester Medical Center	2	MFC N and GoogLeNet	Accuracy = 98.2 % ± 2.4 % and AUC = 94.0 % ± 8.9 %	Not applied
Boulenger et al. [19]	831 BCUS images collected from Peking Union College Hospital	2	CNN with VGG19	AUC = 0.907, F1-score = 79.81 %	4-fold AUC = 0.858 and F1 score = 74.31 %
Gu et al. [20]	14,043 USIs collected from tertiary-care hospitals in China	2	VGG19	Accuracy = 86.40 %, F1-score = 87.2 %, AUC = 0.913, and p-value = 0.5629	Not applied
Chen et al. [22]	880 images (103 normal, 467 malignant and 210 benign)	3	GoogleNet	Training Accuracy = 96.3 % and test accuracy = 93.23 %	Not applied
Shen et al. [23],	US based on clinical data	2	BD-Net	specificity = 91.0 %, sensitivity = 93.8 %, AUC of 0.92, and p-value <0.001	Not applied
Liao et al. [24]	7955 Color Doppler USIs	3	EDL-BC	Accuracy: 97.4 %, F1-Score = 37.0 %, AUC: 0.95, and p-value <0.001	Applied
Wang et al. [25]	769 images	2	ResNet with ASN	ResNet34 v2 Accuracy = 78.11 % and F1-Score = 82 %	Not applied
Taleghamar et al. [26]	QUS multi-parametric imaging	2	Residual Network and RAN	Accuracy = 88 % and AUC: 0.86	Not applied
Fleury & Marcomini [27]	US B-mode images	2	MLP, DT, LDA, RF, and SVM	SVM: Sensitivity = 71.4 % and AUC = 0.840	10-fold applied
Atrey et al. [28]	86 images collected from All India Institute of Medical Sciences, India	2	SVM	Accuracy: 98.84 % and AUC = 0.99	10-fold applied
Wu et al. [29]	US and clinical data of 140 surgically confirmed BC	2	LR	Sensitivity = 86.96 %, AUC = 0.88, and p-value <0.0001	Not applied
Zhang et al. [30]	USIs of 1014 patients	2	Radiomics Nomogram with clinical factors (RNWCF)	Accuracy = 0.846 and AUC = 0.882	5- fold applied
Romeo et al. [32],	USIs	2	RF	Accuracy = 82 %, AUC = 0.90, and p-value: 0.0098	5- fold applied

3. Material and methods

This article suggests a hybrid approach that helps improve survival rates by detecting BC more effectively, as shown in Fig. 1. The online available unbalanced US dataset was used, and before applying the detection process, the images underwent thorough data pre-processing procedures. Three SOTA TL models (MobileNetV2, ResNet50, and VGG16) were combined individually with LSTM and used as feature extractors. These extracted features dataset needed to be more balanced. Due to this, the SMOTETomek method was applied to balance the dataset. Various ML classifiers were used to classify the BC. Finally, the performance was evaluated using multiple metrics and cross-validated with the K-fold method.

3.1. Dataset and image pre-processing

The open-source Kaggle Breast USIs Dataset was chosen for the experiment [34,35]. The dataset comprises 1578 BCE (normal, benign, and malignant) images. It was initially collected from 600 patients between the ages of 25 and 75 in 2018 by Al-Dhabyani et al. [35] and later shared as an open-source resource. The dataset provides essential information about class types, quantities, numbers, percentages, training, validation, and testing details. The dataset has original USIs and masked images, as shown in Fig. 2.

Image pre-processing is crucial in gathering data for DL models. These pre-processing steps simplify complexity, enhance accuracy, and prepare the data for classification algorithms. The BC Ultrasound (BCU) images underwent initial data pre-processing in this study. The pre-processing steps employed on the dataset included resizing the input dimensions to 224 × 224 and the scaling method [36].

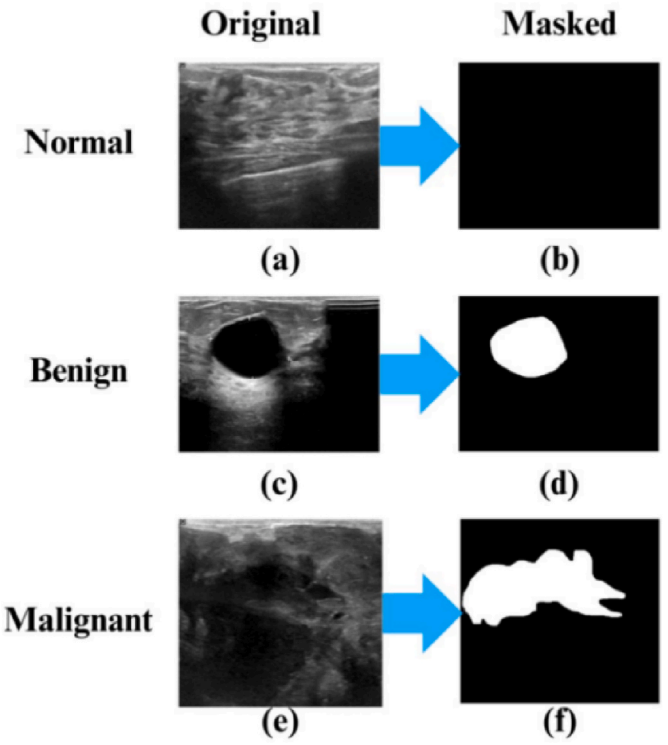

Fig. 2. The Breast USIs dataset images a) original Normal class image, b) masked Normal image, c) Benign class image, d) masked Benign image, e) Malignant class image, and f) masked Malignant image.

3.2. Transfer learning

TL is an ML approach that entails training a model on a particular task and transferring that information to another. It uses learned information from one domain to boost the efficiency of a model on another. TL is beneficial when the target task has insufficient data for training or when the characteristics learned from the source task apply to the target activity. In recent years, a TL model pre-trained on the ImageNet has been employed to classify BC from USIs with excellent outcomes, and it is now a viable option for BC identification. TL can address the issue of inadequate training samples in DL owing to small datasets. In this study, the authors used six TL models: MobileNetV2, ResNet50, VGG16, NASNetMobile [36], Xception [37], and InceptionResNetV2 [38]. The brief description of these models is as follows:

MobilenetV2: MobileNetV2 is a feature extractor that is exceptionally powerful and efficient for mobile device Computer Vision (CV) applications. It outperforms MobileNetV1 in speed, latency, and precision over the whole latency range. The depth-wise separable convolution approach is used in MobileNets [39,40]. MobileNetV2 employs inverted residual blocks with bottlenecking characteristics and two new architectural elements: linear bottlenecks among layers and shortcut connections between bottlenecks. Compared to MobileNetV1, these enhancements offer quicker training and improved precision while minimizing the number of variables and procedures. MobileNetV2's core structure is made up of composite convolutional building blocks

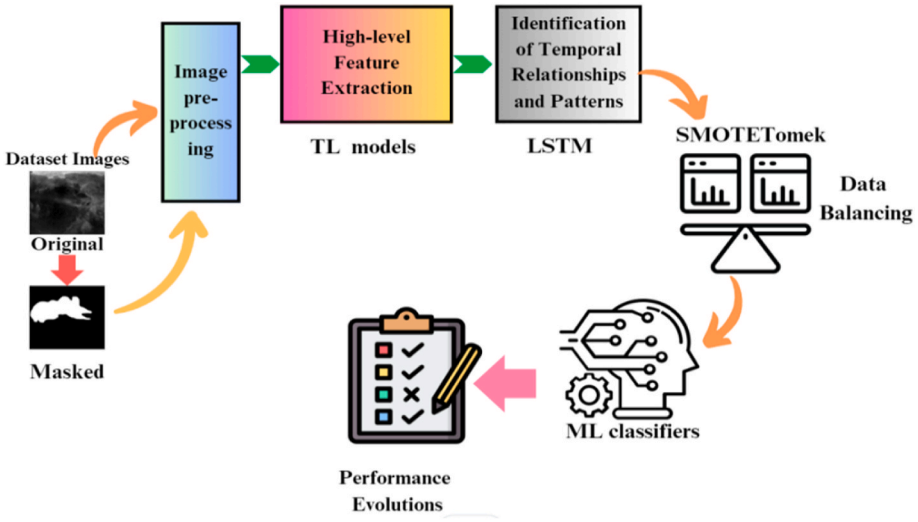

Fig. 1. The flow diagram of proposed TL-LSTM and SMOTETomek with ML classifier framework for detection of BC from US images.

that make effective utilization of depth-wise separable convolution.

ResNet50: ResNet50 is a Deep Convolutional Neural Network (DCNN) developed to handle the vanishing/exploding gradients issue that arises while training DNNs [41]. ResNet adds the idea of shortcut or skip connections, which allow the network to acquire residual mappings that provide feedback to the original input, resulting in residual blocks. This residual method permitted learning to train deeper networks while avoiding optimization issues. ResNet-50's building blocks are designed in a bottleneck fashion, with 1 × 1 convolutions used to decrease the number of parameters and matrix multiplications, allowing for quicker training of each layer. The ResNet50 architecture uses 7 × 7 kernel convolutions, max pooling, and multiple combinations of 3 × 3 and 1 × 1 convolutions repeated numerous times.

VGG16: VGG16 is a DCNN architecture that contributes significantly to CV and image recognition [42]. VGG16 is well-known for its ease of use and constant usage of 3 × 3 filters with a stride of the "same padding". Throughout the design, the convolution and max pooling layers are placed regularly. VGG16 accepts 224 × 224 input tensors with three Red-Green-Blue (RGB) channels. Following the convolutional layers are ReLU activation functions, which add non-linearity to the network's decision-making procedure.

3.3. LSTM

The LSTM network structure was built on top of the Recurrent Neural Network (RNN) to cope with the challenge of long-term sequences. It is exceptional at identifying long-term dependencies, which makes sequence prediction jobs a perfect fit for it. With the addition of feedback connections, LSTM differs from standard neural networks in that it can operate complete data sequences rather of just particular data points. Because of this, it is exceptional at recognizing and forecasting patterns in sequential data. Adding a cell state decides whether past and current information may be added via a gating mechanism, addressing the RNN's "gradient vanishing" and "gradient explosion" difficulties [43]. There are three components in the LSTM network. The term "gates" refers to these three components of an LSTM unit. LSTM unit comprising Forget, Input, and Output gates and a memory cell (LSTM cell), with each neuron having a present state and a hidden layer [44].

Forget gate (f_t): The forget gate regulates whether the previous timestamp must be stored or ignored. An LSTM has a hidden state, where "h_{t-1}" and "h_t" is the hidden state of the previously and currently recorded timestamp, respectively. Furthermore, the cell state of an LSTM is denoted by c_{t-1} for the past timestamp and c_t for the current timestamp, respectively. In this case, long-term memory refers to the cell state and short-term memory to the hidden state. If the x_t is information placed into the memory cell to be trained, and h_t represents the output in every cell, then equation (1) represents the forget gate [45].

$$f_t = \sigma\left(x_t \,.W_f + h_{t-1} \,.V_f\right) \tag{1}$$

where W_f and V_f are the weight matrix related with the input and hidden state, respectively. The sigmoid activation function (σ) adjusts the message weight as it passes through the dot product. As a result, f_t will become a number between 0 and 1 [43,44].

$$c_{t-1} \,.\; f_t = 0 \quad ;\; \text{if } f_t = 0 \tag{2}$$

$$c_{t-1} \,.\; f_t = c_{t-1} \;;\; \text{if } f_t = 1 \tag{3}$$

Input gate (i_t): The cell pursuit to study fresh data from the input to the second gate. The input gate is employed to determine the importance of the fresh data that the input contains [45].

$$i_t = \sigma(x_t \,.W_i + h_{t-1} \,.V_i) \tag{4}$$

where W_i, V_i and h_{t-t} are the weight matrix related with the input, hidden state, and hidden state at the previous timestamp, respectively.

To determine the new information, the function of a concealed state at the time stamp "t-1" and input "x" at the time stamp 't' has to be supplied to the cell state, as shown in the following equation.

$$n_t = \tanh(x_t \,.W_c + h_{t-1} \,.V_c) \tag{5}$$

Here 'tanh' is the activation function, and the fresh data ranges from −1 to 1. Data is added to the cell state if the n_t is positive and subtracted if negative. However, the n_t will not be appended to the cell state immediately.

$$c_t = f_t \,.c_{t-1} + i_t \,.n_t \tag{6}$$

where c_{t-1} is the cell state at the present timestamp.

Output Gate (o_t): Finally, the cell transmits the updated data from the present timestamp to the successive timestamp in the output gate 'o_t' as shown in following equation [45,46].

$$o_t = \sigma\left(x_t \,.W_o + h_{t-1} \,.V_o\right) \tag{7}$$

The resultant number will be from zero to one due to the '' sigmoid function. As illustrated in the following equation, utilize 'o_t' and tanh of the modified cell state to get the present hidden state.

$$h_t = o_t \;.\; \tanh\left(c_t\right) \tag{8}$$

The present output and long-term memory (c_t) determine the concealed state. Use the SoftMax activation on the hidden state h_t to extract the current timestamp output [44].

$$\text{Output} = \text{Softmax}(h_t) \tag{9}$$

3.4. Fusion of TL-LSTM

This work combined the three SOTA TL models, MobileNetV2, ResNet50, and VGG16, with LSTM individually. Table 2 depicts the VGG16-LSTM layer's name, output shape, and the number of parameters. An Input Layer that takes an input tensor with a shape of (None, 224, 224, 3), where the first parameter is the "batch size", the second and third dimensions are the "height" and "width" of the input images, and the fourth parameter is the number of RGB channels. A VGG16 layer that applies the VGG16 to the input tensor and output shape was (None, 7, 7, 512). A Reshape layer that reshapes the output of the VGG16 layer to a shape of (None, 49, 512), which is compatible with LSTM. An LSTM layer that applies a LSTM RNN to the reshaped input and output shape was (None, 49, 100). A flattened layer flattens the output of the LSTM layer to a shape of (None, 4900). Two dense layers were used with 25 and 3 neurons, respectively. The output shape of dense1 was (None, 25), and dense2 was (None, 3). The flattened layer features were applied to the SMOTomek for balancing.

3.5. Synthetic Minority Oversampling Technique (SMOTE)

Because the data for every category is uneven, a classification bias may favour the majority category and under-sample the minority category. Chawla et al. [47] advised tackling the data imbalance issue using

Table 2
VGG16-LSTM models details with layer type, output shape, and number of parameters.

Layer (type)	Output Shape	Param #
input (InputLayer)	[(None, 224, 224, 3)]	0
vgg16 (Functional)	(None, 7, 7, 512)	14714688
reshape (Reshape)	(None, 49, 512)	0
lstm (LSTM)	(None, 49, 100)	245200
flatten (Flatten)	(None, 4900)	0
dense1 (Dense)	(None, 25)	122525
dense2 (Dense)	(None, 3)	78
Total params: 15082491 (57.54 MB)		

the SMOTE technique. Random interpolation is performed among the sample feature space for every target group and its closest neighbor to synthesize a new sample. This can assist the classifier in boosting its generalization ability by enhancing the number of minority categories [48]. The SMOTE oversampling and Tomek Links under-sampling methods are combined in SMOTomek. The SMOTE generates artificial information for the minority class, whereas Tomek Links eliminates information about the majority class, recognized as Tomek Links. The SMOTE phase starts by finding the number of nearest neighbours (k) and subsequently employing the Euclidean Distance (ED), computing the distance that is shortest among the random data taken from the minority class (x_{ci}) and the data of the k-nearest neighbours (x_{ki}). In addition, depending on the shortest distance, artificial sample data (x_{si}) for the minority class are created using the following equation [48]:

$$x_{si} = x_{ci} + r(x_{ci} - x_{ki}) \tag{10}$$

The procedure is terminated once the information for every category is balanced [49]. The initial stage of Tomek Links is to select a pair of examples with the shortest ED from the k-nearest neighbours, with everyone belonging to a distinct category (x_g, x_h). The (x_g, x_h) is a Tomek Link; if no sample x_k meets the subsequent ED requirements.

$$d(x_g, x_k) < d(x_g, x_h) \text{ or } d(x_h, x_k) < d(x_g, x_h) \tag{11}$$

4. Results

The analysis was carried out on Google Colab PRO with the Python3 Google compute engine backend (GPU) with 25.45 GHz RAM and 166.75 GB of disc space, offering programmers a robust environment for constructing and training DL models. Python and popular libraries like TensorFlow, Keras, Scikit-learn, Pandas, NumPy, and Matplotlib make model creation and data analysis easier. The "Adam" (learning rate = 0.001, beta1 = 0.9, and beta 2 = 0.999) optimizer and the "categorical cross-entropy" loss function were used. It changes the learning rates of individual parameters based on their past gradients, making it suited for a wide range of DL applications. The investigators chose a "batch size" of "32" to train the DL models while adequately using the most computer power. The "batch size = 32" was a good compromise between effectiveness and memory constraints. The dataset set was split in the 90:10 ratio for training and test sets.

4.1. Performance measures

The performance measures used to check the suggested system's effectiveness include accuracy, recall, precision, F1-score, MCC, Cohen Kappa Score (K) [50], AUC-ROC curve, and cross-validation with the K-fold technique [51]. These performance metrics are critical for assessing the usefulness and reliability of DL models.

$$\text{Accuracy (A)} = \frac{\text{True Positive } (TP_{bc}) + \text{True Negative } (TN_{bc})}{\text{True Positive } (TP_{bc}) + \text{False Positive } (FP_{bc}) + \text{True Negative } (TN_{bc}) + \text{False Negative } (FN_{bc})} \tag{12}$$

$$\text{Recall (R)} = \frac{TP_{bc}}{TP_{bc} + FN_{bc}} \tag{13}$$

$$\text{Precision (P)} = \frac{TP_{bc}}{TP_{bc} + FP_{bc}} \tag{14}$$

$$\text{F1} - \text{Score (F1)} = \frac{2}{\left(\frac{1}{R_{bc}}\right) + \left(\frac{1}{P_{bc}}\right)} \tag{15}$$

$$\text{MCC} = \frac{((TP_{bc} \times TN_{bc} - FP_{bc} \times FN_{bc})}{\sqrt{((TP_{bc} + FP_{bc}) \times (TP_{bc} + FN_{bc}) \times (TN_{bc} + FP_{bc}) \times (TN_{bc} + FN_{bc}))}} \tag{16}$$

$$\text{Cohen's Kappa (K)} = \frac{\text{Actual agreement } (P_o) - \text{Projected agreement } (P_e)}{1 - \text{Projected agreement } (P_e)} \tag{17}$$

Landis and Koch present a technique for characterizing values: 0 for no agreement, 0–0.20 for modest agreement, 0.21–0.40 for reasonable agreement, 0.41–0.60 for moderate agreement, 0.61–0.80 for significant agreement and 0.81–1 means practically perfect agreement [52].

4.2. Performance of DCNN-LSTM models

This section involves collecting features from dataset images via MobileNet-LSTM, ResNet50-LSTM, and VGG16-LSTM, sending these extracted features to numerous classifiers, and assessing the efficacy via different metrics, as shown in Table 3. The dataset set was split in the 90:10 ratio for training and test sets. Accuracy (A_{bc}), precision (P_{bc}), recall (R_{bc}), F1 score ($F1_{bc}$), MCC (M_{bc}), Cohen's Kappa (K_{bc}), and Macro Average AUC (MAAUC) are essential metrics utilized to estimate the classifier's usefulness. The accuracy is the proportion of adequately categorized instances divided by the total number. On the other hand, P_{bc} sets the model's capacity to accurately identify positive cases among all positive cases. R_{bc} measures the model's ability to recognize entirely positive issues among all positive outcomes. The F1 score provides an appropriate measure of precision and recall when handling imbalanced datasets. MCC is a performance metric considering TP_{bc}, TN_{bc}, FP_{bc}, and FN_{bc}. It ranges from −1 to +1, with +1 indicating a perfect classifier, 0 indicating a random classifier, and −1 indicating a completely inaccurate classifier. K_{bc} measures the degree of agreement between predicted and true labels while considering the probability of chance agreement. It is advantageous when handling datasets that have imbalanced class distributions.

The SVCR classifier achieved A_{bc}, P_{bc}, R_{bc}, F_{bc}, M_{bc}, K_{bc}, and MAAUC of 94.0 %, 94.0 %, 92.0 %, 93.0 %, 89.2 %, 89.0 %, and 98.0 % with MobileNet-LSTM. A P_{bc} of 94.0 % indicates that 94.0 % of the predicted positives were correct, whereas an R_{bc} of 92.0 % suggests that the SVCR correctly detected 92.0 % of the positive samples. An F1-score of 93.0 % shows a balanced trade-off between P_{bc} and R_{bc}. The M_{bc} of 89.2 % indicates that the predictions of the SVCR model precisely correspond to the actual labels, showing its accuracy in recognizing BC from USIs. A K_{bc} of 89.0 % offers a high level of agreement. MAAUC is a metric that assesses the performance of a classifier over several classes, taking into account the Receiver Operating Characteristic (ROC) curve for every category. MAAUC of 98.0 % shows greater efficacy in multi-class categorization. Furthermore, the authors utilized K-fold cross-validation (K = 10) to assess the classifiers' validity. K-fold cross-validation is an approach that includes partitioning the dataset into K subgroups and training the model K times. This strategy assures that the training-test split does not affect the model's effectiveness. The SVCR had the highest scores in K-fold accuracy (KA_{bc}), precision (KP_{bc}), recall (KR_{bc}), F1

Table 3
Performance of the classifiers with MobileNet-LSTM, ResNet50-LSTM, and VGG16-LSTM extracted features.

Methods	Classifiers	A_{bc}	P_{bc}	R_{bc}	F_{bc}	M_{bc}	K_{bc}	MA-AUC	KA_{bc}	KR_{bc}	KR_{bc}	KF_{bc}	KM_{bc}	KK_{bc}
MobileNetV2-LSTM features	LR-N	89.0 %	88.0 %	88.0 %	88.0 %	80.1 %	80.3 %	98.0 %	82.9 %	80.2 %	82.4 %	80.9 %	70.6 %	70.2 %
	LR-L	89.0 %	89.0 %	88.0 %	88.0 %	81.5 %	81.3 %	98.0 %	82.4 %	80.4 %	81.3 %	80.5 %	69.9 %	69.7 %
	SVCR	**94.0 %**	**94.0 %**	**92.0 %**	**93.0 %**	**89.2 %**	**89.0 %**	**98.0 %**	**84.8 %**	**80.7 %**	**86.9 %**	**82.8 %**	**73.6 %**	**72.7 %**
	SVCL	90.0 %	89.0 %	89.0 %	89.0 %	82.6 %	82.6 %	97.0 %	82.8 %	80.5 %	82.1 %	80.9 %	70.4 %	70.2 %
	DTC	78.0 %	78.0 %	74.0 %	76.0 %	62.1 %	61.5 %	84.0 %	72.2 %	69.4 %	69.9 %	69.4 %	52.7 %	52.5 %
	KNC	91.0 %	92.0 %	90.0 %	90.0 %	85.1 %	84.6 %	93.0 %	82.3 %	80.7 %	80.6 %	80.3 %	70.1 %	69.7 %
	RFC	89.0 %	92.0 %	84.0 %	87.0 %	80.2 %	79.2 %	96.0 %	81.2 %	72.1 %	88.9 %	76.9 %	67.5 %	64.4 %
	GNB	76.0 %	74.0 %	79.0 %	74.0 %	63.9 %	62.3 %	87.0 %	68.8 %	70.8 %	69.1 %	66.2 %	53.6 %	51.1 %
ResNet50-LSTM features	**LR-N**	**92.0 %**	**93.0 %**	**90.0 %**	**91.0 %**	**85.8 %**	**85.5 %**	**99.0 %**	**87.2 %**	**85.1 %**	**87.7 %**	**86.1 %**	**77.8 %**	**77.6 %**
	LR-L	90.0 %	88.0 %	91.0 %	90.0 %	83.2 %	83.1 %	97.0 %	83.9 %	84.7 %	82.1 %	83.0 %	73.7 %	73.2 %
	SVCR	92.0 %	93.0 %	92.0 %	92.0 %	87.0 %	86.7 %	99.0 %	86.7 %	82.4 %	90.2 %	85.2 %	76.9 %	75.9 %
	SVCL	91.0 %	89.0 %	92.0 %	90.0 %	85.2 %	85.0 %	97.0 %	87.0 %	85.1 %	87.4 %	86.0 %	77.4 %	77.2 %
	DTC	78.0 %	76.0 %	80.0 %	78.0 %	94.3 %	94.1 %	84.0 %	74.6 %	72.0 %	71.6 %	71.7 %	56.6 %	56.4 %
	KNC	91.0 %	91.0 %	90.0 %	90.0 %	83.8 %	83.8 %	93.0 %	81.1 %	80.1 %	80.0 %	79.7 %	68.3 %	67.9 %
	RFC	90.0 %	93.0 %	87.0 %	90.0 %	82.6 %	81.8 %	97.0 %	84.6 %	77.7 %	90.6 %	81.8 %	73.5 %	71.3 %
	GNB	82.0 %	79.0 %	78.0 %	78.0 %	68.9 %	68.7 %	91.0 %	74.7 %	73.7 %	73.8 %	73.1 %	58.3 %	57.7 %
VGG16-LSTM features	**LR-N**	**94.0 %**	**94.0 %**	**94.0 %**	**94.0 %**	**89.1 %**	**89.1 %**	**99.0 %**	**88.7 %**	**86.7 %**	**89.2 %**	**87.6 %**	**80.5 %**	**80.2 %**
	LR-L	93.0 %	90.0 %	95.0 %	92.0 %	88.7 %	88.4 %	98.0 %	87.3 %	87.1 %	85.8 %	86.0 %	78.9 %	78.6 %
	SVCR	92.0 %	93.0 %	92.0 %	92.0 %	97.1 %	86.8 %	99.0 %	88.8 %	85.1 %	91.5 %	87.4 %	80.7 %	79.9 %
	SVCL	93.0 %	93.0 %	93.0 %	93.0 %	88.1 %	88.0 %	98.0 %	89.1 %	87.3 %	89.4 %	88.1 %	81.3 %	81.0 %
	DTC	82.0 %	81.0 %	78.0 %	80.0 %	67.9 %	67.6 %	86.0 %	77.2 %	74.9 %	75.5 %	74.8 %	61.2 %	60.9 %
	KNC	92.0 %	89.0 %	93.0 %	90.0 %	86.7 %	83.2 %	94.0 %	78.7 %	80.7 %	76.1 %	76.1 %	67.5 %	65.9 %
	RFC	89.0 %	93.0 %	85.0 %	88.0 %	80.8 %	79.3 %	98.0 %	85.1 %	78.8 %	90.8 %	82.7 %	74.4 %	72.4 %
	GNB	85.0 %	84.0 %	80.0 %	81.0 %	74.6 %	74.5 %	94.0 %	80.4 %	78.5 %	78.6 %	78.3 %	66.8 %	66.5 %

score (KF_{bc}), MCC (KM_{bc}), and Cohen's Kappa (KK_{bc}) with values of 84.8 %, 80.7 %, 86.9 %, 82.8 %, 73.6 %, and 72.7 %, respectively. The outcome is less favourable than the single-fold measurements, but it is appropriate, indicating that the SVCR model operates adequately across several information folds.

The SVCR classifier achieved A_{bc}, P_{bc}, R_{bc}, F_{bc}, M_{bc}, K_{bc}, and MAAUC of 92.0 %, 93.0 %, 92.0 %, 92.0 %, 87.0 %, 86.7 %, and 99.0 % with ResNet50-LSTM, while the highest KA_{bc}, KP_{bc}, KR_{bc}, KF_{bc}, KM_{bc}, and KK_{bc} had values of 86.7 %, 82.4 %, 90.2 %, 85.2 %, 76.9 %, and 75.9 %, respectively. On the other hand, the LR-N classifier achieved A_{bc}, P_{bc}, R_{bc}, F_{bc}, M_{bc}, K_{bc}, and MAAUC of 94.0 %, 94.0 %, 94.0 %, 94.0 %, 89.1 %, 89.1 %, and 99.0 % with VGG16-LSTM, while the highest KA_{bc}, KP_{bc}, KR_{bc}, KF_{bc}, KM_{bc}, and KK_{bc} had values of 88.7 %, 86.7 %, 89.2 %, 87.6 %, 80.5 %, and 80.2 %, respectively.

The highest performance was achieved with VGG16-LSTM among the three methods. While the models are generally effective, the authors acknowledge that specific performance measures might be improved, notably the 10-fold KM_{bc} and KK_{bc} values. The authors consider future adjustments to enhance performance.

4.3. Performance of DCNN-LSTM with SMOTETomek

The dataset used in this study has 891 benign, 421 malignant, and 266 normal images. The dataset needs to be more balanced. Due to this, the authors used the SMOTETomek data sampling method to balance the dataset. First, the DCNN-LSTM models extracted the features from the images, and then the SMOTETomek was applied to these extracted features. The dataset size changes from 1578 × 4900 to 2671 × 4900. The SMOTETomek performed features were employed to the classifiers to predict the benign, malignant, and normal. Table 4 depicts the performance of classifiers.

The SVCR achieved A_{bc}, P_{bc}, R_{bc}, F_{bc}, M_{bc}, K_{bc}, and MAAUC of 99.0 %, 99.0 %, 99.0 %, 98.0 %, 97.8 %, 97.7 %, and 100.0 %, respectively with MobileNetV2-LSTM-SMOTETomek, while the highest KA_{bc}, KP_{bc}, KR_{bc}, KF_{bc}, KM_{bc}, and KK_{bc} had values of 95.7 %, 95.7 %, 95.8 %, 95.7 %, 93.6 %, and 93.5 %, respectively. The performance of RFC was also improved with A_{bc}, P_{bc}, R_{bc}, F_{bc}, M_{bc}, K_{bc}, and MAAUC of 99.0 %, 99.0 %, 99.0 %, 99.0 %, 98.3 %, 98.3 %, and 100.0 %, respectively, while KA_{bc}, KP_{bc}, KR_{bc}, KF_{bc}, KM_{bc}, and KK_{bc} had values of 94.1 %, 94.1 %, 94.5 %, 94.1 %, 91.3 %, and 91.1 %, respectively. The remaining

Table 4
Performance of the classifiers with balanced features of MobileNet-LSTM-SMOTETomek, ResNet50-LSTM-SMOTETomek, and VGG16-LSTM-SMOTETomek.

Methods	Classifiers	A_{bc}	P_{bc}	R_{bc}	F_{bc}	M_{bc}	K_{bc}	MA-AUC	KA_{bc}	KR_{bc}	KR_{bc}	KF_{bc}	KM_{bc}	KK_{bc}
MobileNetV2-LSTM features	LR-N	97.0 %	97.0 %	97.0 %	97.0 %	95.6 %	95.5 %	100.0 %	94.2 %	94.2 %	94.4 %	94.1 %	91.4 %	91.2 %
	LR-L	98.0 %	98.0 %	98.0 %	98.0 %	96.7 %	96.6 %	100.0 %	94.1 %	94.1 %	94.3 %	94.1 %	91.3 %	91.2 %
	SVCR	**99.0 %**	**99.0 %**	**99.0 %**	**98.0 %**	**97.8 %**	**97.7 %**	**100.0 %**	**95.7 %**	**95.7 %**	**95.8 %**	**95.7 %**	**93.6 %**	**93.5 %**
	SVCL	98.0 %	98.0 %	98.0 %	98.0 %	96.7 %	96.6 %	100.0 %	94.1 %	94.1 %	94.3 %	94.0 %	91.3 %	91.1 %
	DTC	86.0 %	86.0 %	86.0 %	86.0 %	79.7 %	79.2 %	90.0 %	83.3 %	83.3 %	83.5 %	83.2 %	75.2 %	75.1 %
	KNC	95.0 %	96.0 %	95.0 %	95.0 %	92.9 %	92.7 %	96.0 %	89.2 %	89.3 %	90.6 %	88.6 %	84.9 %	83.9 %
	RFC	99.0 %	99.0 %	99.0 %	99.0 %	98.3 %	98.3 %	100.0 %	94.1 %	94.1 %	94.5 %	94.1 %	91.3 %	91.1 %
	GNB	77.0 %	80.0 %	77.0 %	77.0 %	68.1 %	68.8 %	88.0 %	74.4 %	74.4 %	76.6 %	74.2 %	62.8 %	61.7 %
ResNet50-LSTM features	LR-N	97.0 %	98.0 %	97.0 %	97.0 %	96.1 %	96.0 %	100.0 %	95.8 %	95.8 %	96.0 %	95.7 %	93.8 %	93.7 %
	LR-L	98.0 %	98.0 %	98.0 %	98.0 %	96.7 %	96.6 %	100.0 %	95.7 %	95.7 %	95.9 %	95.7 %	93.7 %	93.6 %
	SVCR	99.0 %	99.0 %	99.0 %	99.0 %	98.8 %	98.8 %	100.0 %	96.0 %	96.0 %	96.4 %	96.0 %	94.1 %	93.9 %
	SVCL	**97.0 %**	**98.0 %**	**97.0 %**	**97.0 %**	**96.1 %**	**96.1 %**	**99.0 %**	**96.0 %**	**96.0 %**	**96.3 %**	**96.0 %**	**94.2 %**	**94.1 %**
	DTC	90.0 %	90.0 %	90.0 %	89.0 %	84.5 %	84.3 %	92.0 %	83.3 %	83.3 %	83.7 %	83.1 %	75.2 %	74.9 %
	KNC	95.0 %	95.0 %	95.0 %	95.0 %	92.8 %	92.7 %	96.0 %	88.6 %	88.6 %	90.3 %	87.8 %	84.1 %	82.9 %
	RFC	97.0 %	97.0 %	97.0 %	97.0 %	96.1 %	96.1 %	100.0 %	94.2 %	94.2 %	94.7 %	94.2 %	91.5 %	91.3 %
	GNB	90.0 %	91.0 %	90.0 %	90.0 %	84.8 %	84.2 %	95.0 %	81.8 %	81.9 %	83.9 %	81.9 %	73.6 %	72.7 %
VGG16-LSTM features	LR-N	99.0 %	99.0 %	99.0 %	99.0 %	98.9 %	98.9 %	100.0 %	96.0 %	96.0 %	96.3 %	96.0 %	94.1 %	94.0 %
	LR-L	**99.0 %**	**99.0 %**	**99.0 %**	**99.0 %**	**98.9 %**	**98.9 %**	**100.0 %**	**96.1 %**	**96.1 %**	**96.3 %**	**96.0 %**	**94.2 %**	**94.1 %**
	SVCR	99.0 %	99.0 %	99.0 %	99.0 %	97.8 %	97.7 %	100.0 %	96.0 %	96.0 %	96.4 %	96.0 %	94.2 %	94.0 %
	SVCL	**99.0 %**	**99.0 %**	**99.0 %**	**99.0 %**	**98.9 %**	**98.9 %**	**100.0 %**	**96.1 %**	**96.1 %**	**96.3 %**	**96.0 %**	**94.2 %**	**94.1 %**
	DTC	94.0 %	94.0 %	94.0 %	94.0 %	91.2 %	92.0 %	95.0 %	86.7 %	86.7 %	87.1 %	86.7 %	80.3 %	80.1 %
	KNC	95.0 %	96.0 %	95.0 %	95.0 %	92.9 %	92.7 %	96.0 %	88.6 %	88.5 %	90.7 %	88.0 %	84.1 %	82.8 %
	RFC	98.0 %	99.0 %	99.0 %	99.0 %	97.8 %	97.7 %	100.0 %	94.1 %	94.1 %	94.9 %	94.1 %	91.5 %	91.1 %
	GNB	93.0 %	94.0 %	93.0 %	93.0 %	90.0 %	89.8 %	97.0 %	83.6 %	83.6 %	86.2 %	83.6 %	76.5 %	75.5 %

classifieris performance also improved compared to the without applying the SMOTETomek method.

On the other hand, the SVCR achieved A_{bc}, P_{bc}, R_{bc}, F_{bc}, M_{bc}, K_{bc}, and MAAUC of 99.0 %, 99.0 %, 99.0 %, 99.0 %, 98.8 %, 98.8 %, and 100.0 % with ResNet50-LSTM-SMOTETomek, while the highest KA_{bc}, KP_{bc}, KR_{bc}, KF_{bc}, KM_{bc}, and KK_{bc} had values of 96.0 %, 96.0 %, 96.4 %, 96.0 %, 94.1 %, and 93.9 %, respectively. The remaining classifiers performance also improved. The LR-N, LR-L, SVCL and SVCR performed similarly with VGG16-LSTM-SMOTETomek method. All these four classifier's achieved A_{bc}, P_{bc}, R_{bc}, F_{bc}, M_{bc}, K_{bc}, and MAAUC of 99.0 %, 99.0 %, 99.0 %, 99.0 %, 98.9 %, 98.9 %, and 100.0 %, respectively, while the highest KA_{bc}, KP_{bc}, KR_{bc}, KF_{bc}, KM_{bc}, and KK_{bc} had values of 96.1 %, 96.1 %, 96.3 %, 96.0 %, 94.2 %, and 94.1 %, respectively, achieved by LR-L and SVCL models. The remaining classifiers performance also improved. The VGG16-LSTM-SMOTETomek method improves the performance compared to the MobileNetV2 and ResNet50-LSTM-SMOTETomek methods.

From the above two scenarios, the SVCR, SVCL, RFC, LR-N and LR-L models perform better than the other classifiers and achieve an AUC of 1. The AUC-ROC performance of these classifiers is shown in Fig. 3.

4.4. LIME

LIME is a model-independent method that aids in analyzing the model by varying the input of data points and observing how it forecasts variations. LIME alters one example of data by changing the feature values and then examines the effect on the output. LIME generates a list of interpretations that indicate the importance of each attribute to the forecast for a data sample. LIME enables local interpretability while determining which feature modifications influence the forecast most. LIME is model-agnostic, relies on primary and comprehensible ideas, and requires little effort. Fig. 4 shows the original dataset images with LIME-applied images.

4.5. Gradient-weighted Class Activation Mapping (Grad-CAM)

Grad-CAM is an approach for visualizing DCNN explanations by creating visual explanations for judgements made by a wide range of models. Grad-CAM generates a heatmap highlighting the parts of an input picture most significant for a specific class. This heatmap may be used to understand why the model predicted what it did and to detect

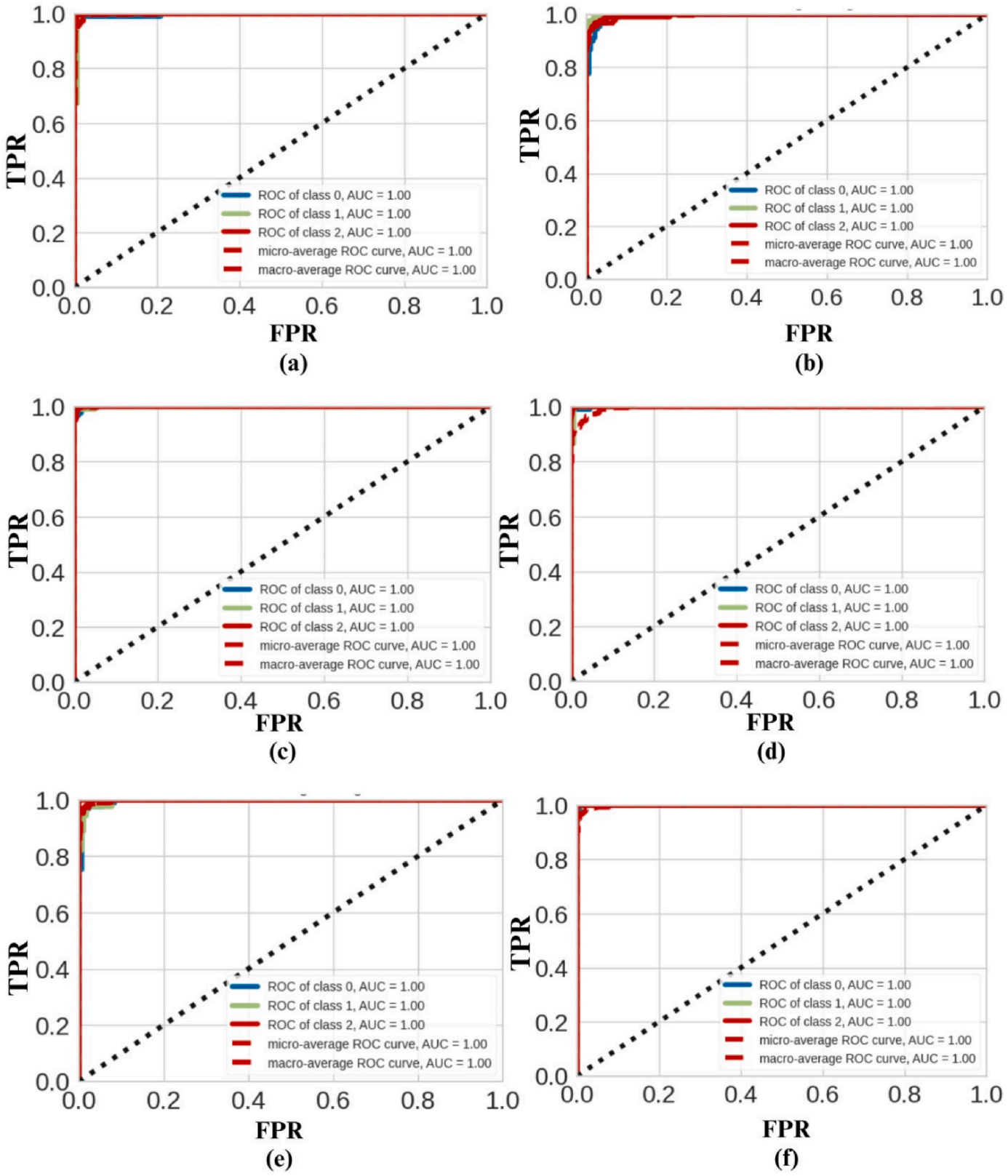

Fig. 3. AUC-ROC curve performance of a) MobileNet -SVCR, b) MobileNet –RFC, c) ResNet -SVCR, d) VGG -LR-N, e) VGG -LR-L, and f) VGG-SVCL (class-0 for "Benign", class-1 for "Malignant", and class-2 for "Normal").

potential model biases. Fig. 5 shows the Grad-CAM plot of MobileNet, ResNet50, and VGG16 models. Fig. 5 shows that the VGG16 model identifies the area of interest most efficiently compared to the other model.

4.6. Normal approximation interval (NAI)

The NAI is the statistical approach for calculating CIs for a population parameter with a normal distribution from a single training-test

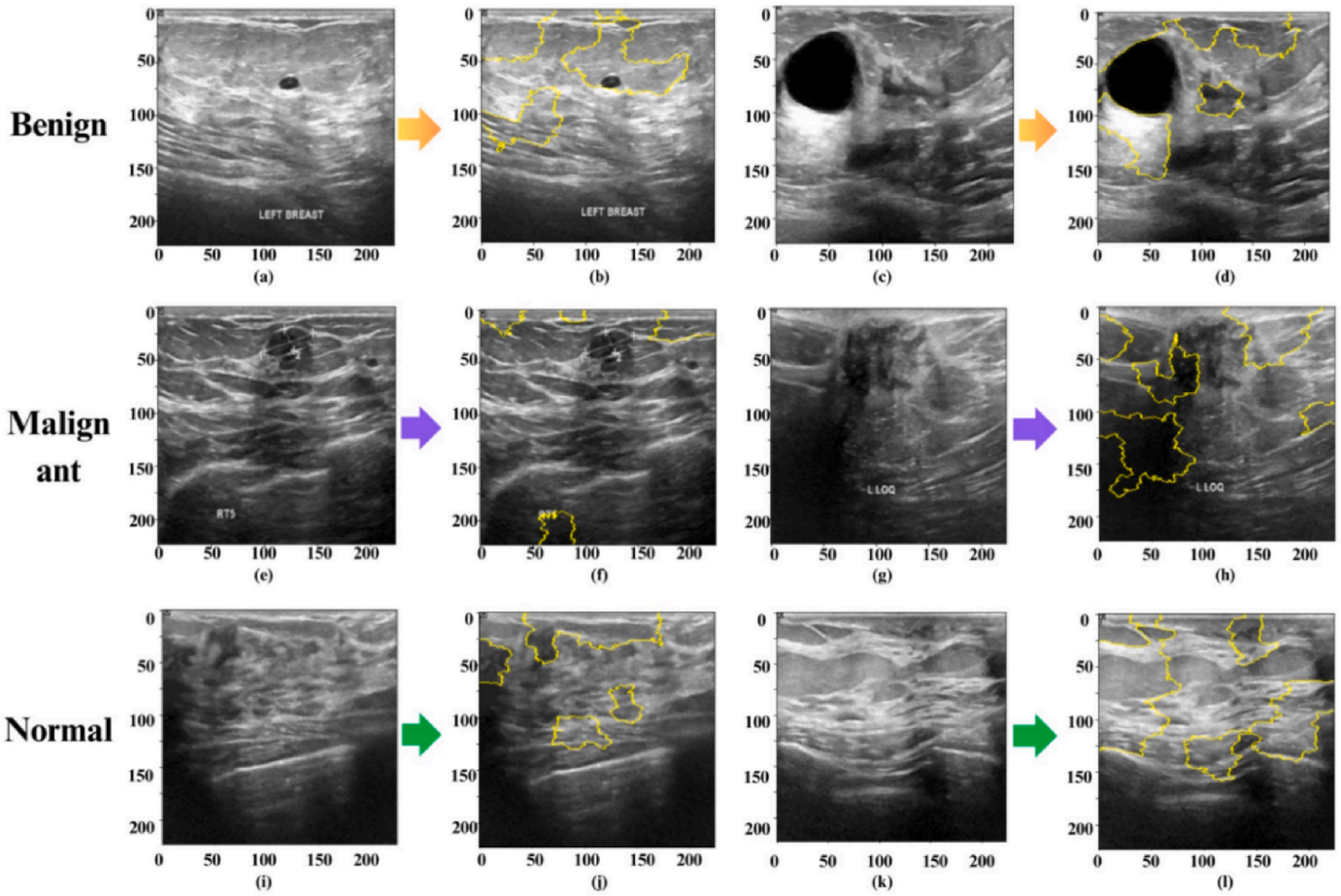

Fig. 4. Original and LIME applied breast cancer USIs: a) original benign image, b) LIME-applied benign image, c) original benign image, d) LIME-applied benign image, e) original malignant image, f) LIME-applied malignant image, g) original malignant image, h) LIME-applied malignant image, i) original normal image, j) LIME-applied normal image, k) original normal image, l) LIME-applied normal image.

division [53,54]. NAI is especially appealing in DL, wherein model training is costly. It is beneficial when dealing with large datasets and includes approximating the sample statistic's distribution using a normal distribution. The NAI is based on the Central Limit Theorem, which affirms that as sample size hikes, the sampling distribution of the sample mean continues to adhere to a normal distribution, independent of the underlying distribution of the population. This theorem is the foundation for inferring population parameters using the normal distribution [54]. Fig. 6 visually depicts the CIs for each classifier, demonstrating their variety and dispersion. This visualization compares the efficacy and dependability of various classifiers. Table 5 shows a tabular overview of the lower and upper bounds of the CIs, as well as the mean value determined for several classifiers. Fig. 6 and Table 5 show that the LR-N and SVCR classifiers have the most significant values for the lower CI, higher CI, and mean among the classifiers evaluated. The LR-N classifier's lower, higher, and mean CIs were 96.50 %, 99.75 %, and 98.13 %, respectively. The narrowness of the CIs indicates the robustness and accuracy of these classifiers in delivering forecasts that are not only consistent but also have high degrees of certainty.

4.7. Bootstrapping method

Understanding the method's assumptions and restrictions is critical, especially when the dataset is small or the population distribution deviates significantly from normalcy. The bootstrapping or non-parametric techniques may be more suited for reliable CI estimates. Moreover, if one projection (accuracy) is from one test set, we must hypothesize how accuracy is distributed. Let us assume that the values for accuracy are regularly distributed, so we have to examine the 95 % accuracy range. In such situations, the bootstrap method helps to assess models. In the bootstrap method, the model trained on training folds and evaluated performance on the test set from every iteration. Furthermore, Bouthillier et al. [48] discovered that an out-of-bag bootstrap approach can increase effectiveness and estimate reliability.

Table 6 is a tabular representation of the 95 % lower confidence interval (LCI) and higher confidence interval (HCI) and the bootstrap mean for all classifiers used in the study. This table presents a structured overview of major statistical parameters characterizing the classifiers' effectiveness and dependability regarding prediction results. The LCI and HCI are essential components in statistical inference because they provide a range within which the actual population parameter is predicted to lie with a certain level of certainty. The 95 % CI in this situation indicates a 95 % chance that the real parameter value falls inside the estimated interval. On the other hand, the bootstrap mean is generated from several bootstrap samples created by repetitive resampling of the original dataset. The efficiency metric of the classifier is generated for every sample, and the mean of those metrics throughout all samples indicates the classifier's overall predicted reliability. According to Table 6, the LR-N, LR-L, SVC-R, and SVCL models have identical values for lower and higher CIs and the bootstrap mean. This shows that their prediction performance is consistent and comparable. The SVCR model, on the other hand, stands out with the most significant values for both the lower and upper CIs and the bootstrap mean. The SVCR model, in particular, has a 95 % LCI of 93.81 %, an HCI of 96.00 %, and a bootstrap mean of 94.90 %. According to these data, the SVCR model has better predictive performance and is projected to retain high accuracy within

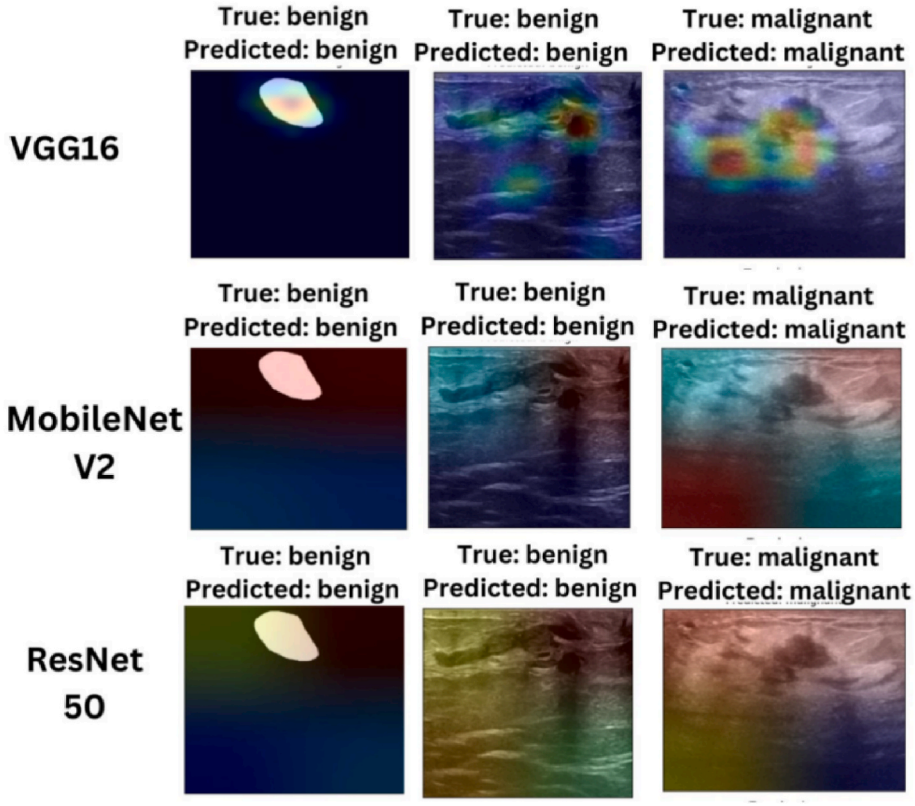

Fig. 5. Grad-CAM visualization plot of VGG16, MobileNetV2, and ResNet50 models.

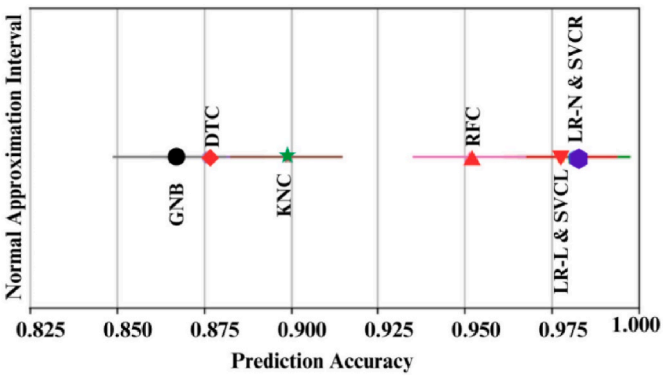

Fig. 6. Mean of LCI and UCI of all classifiers with VGG16-LSTM-SMOTETomek using NAI method.

Table 5
LCI, UCI and mean of various classifiers with VGG16-LSTM-SMOTETomek.

Models	LCI	UCI	Mean
LR-N	96.50 %	99.75 %	98.13 %
LR-L	96.13 %	99.38 %	97.76 %
SVCR	96.50 %	99.75 %	98.13 %
SVCL	96.13 %	99.38 %	97.76 %
DTC	86.26 %	89.27 %	87.77 %
KNC	88.26 %	91.51 %	89.89 %
RFC	93.51 %	96.76 %	95.14 %
GNB	84.89 %	88.14 %	86.52 %

the defined confidence ranges. Fig. 7 depicts the 95 % LCI, 95 % HCI, and bootstrap mean for the different classifiers in line with Table 6. This graphical depiction contributes to a clear and intuitive comprehension

Table 6
LCI, UCI and bootstrap mean of various classifiers with VGG16-LSTM-SMOTETomek using bootstrap method.

Models	LCI	UCI	Bootstrap mean
LR-N	93.08 %	95.52 %	94.30 %
LR-L	93.12 %	95.86 %	94.49 %
SVCR	93.81 %	96.00 %	94.90 %
SVCL	93.40 %	95.41 %	94.41 %
DTC	81.33 %	86.99 %	84.16 %
KNC	83.15 %	88.15 %	85.65 %
RFC	91.63 %	95.41 %	93.52 %
GNB	80.44 %	85.68 %	83.06 %

of the variability and distribution of these metrics among classifiers.

4.8. Comparison of proposed methods with states of art (SOTA) models

In this part, the performance of the six SOTA models: Inception-ResNetV2 [56], Xception [57], NASNetMobile, MobileNetV2, ResNet50, and VGG16, were assessed about the suggested technique (VGG16-LSTM -SMOTETomek-LR_L). It is critical to compare a suggested approach to SOTA models to assess its efficacy and uniqueness. This aids in benchmarking, reaching the effectiveness of existing solutions, and identifying strengths and flaws, allowing for future refinement and development. This demonstrates that the suggested strategy outperforms or matches SOTA models, demonstrating its innovation and contribution to the area. The 10-fold cross-validation approach was employed to test the effectiveness of all SOTA models and suggested methods, and the outcomes are listed in Table 7.

From Table 7, InceptionResNetV2 achieved the best 10-fold average performance metrics KA_{bc}, KP_{bc}, KR_{bc}, KF_{bc}, KM_{bc}, and KK_{bc} of 94.5 %, 88.6 %, 88.7 %, 88.4 %, 81.7 %, and 81.4 %, respectively, among the SOTA models. NASNetMobile followed closely with 89.8 %, 89.1 %,

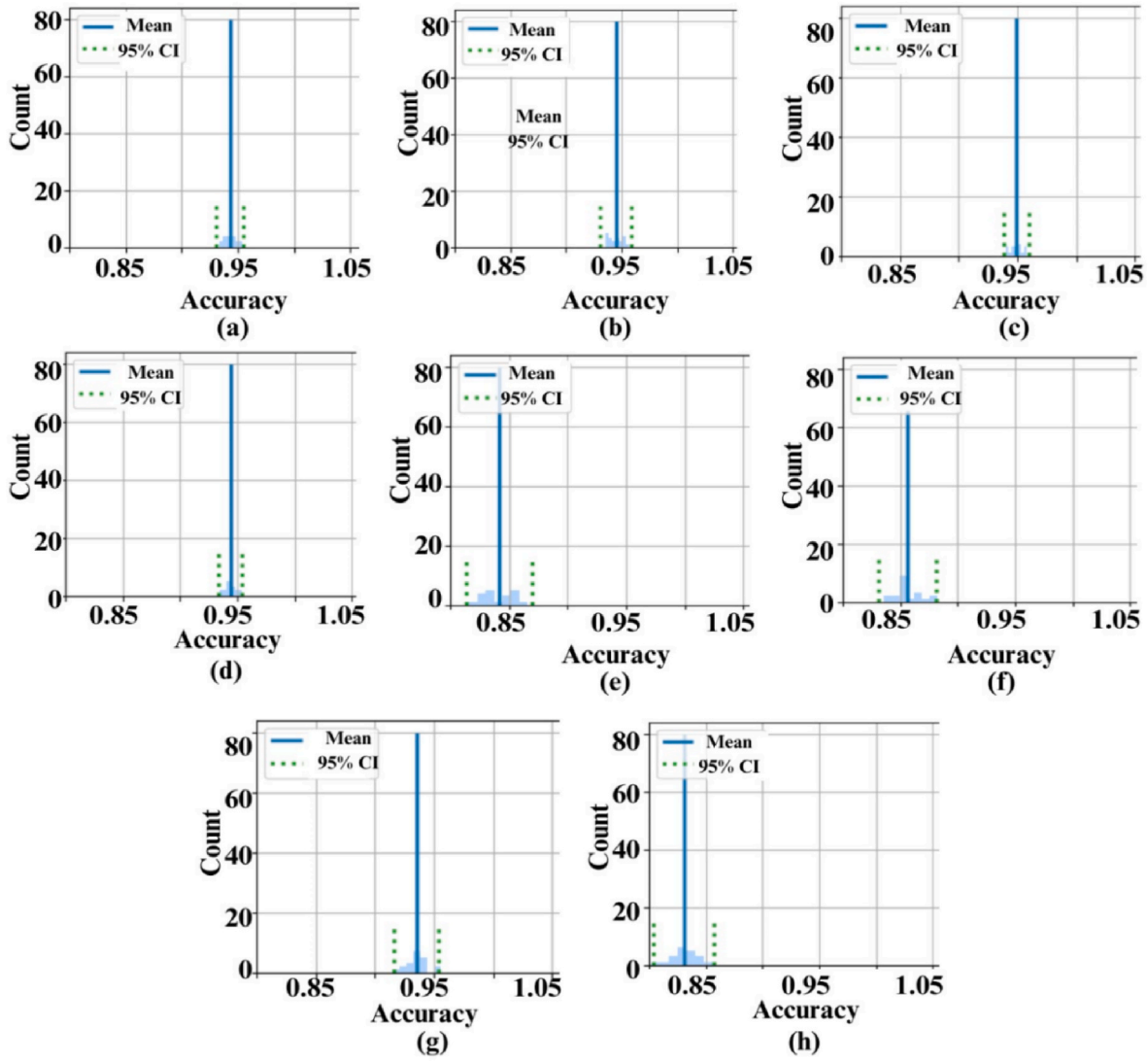

Fig. 7. 95 % LCI, and UCI with mean of various classifiers using bootstrap method update: a) LR-N, b) LR-L, c) SVCR, d) SVCL, e) DTC, f) KNC, g) RFC, and h) GNB (vertical blue line indicates mean of LCI and UCI and two vertical dotted lines indicates LCI and UCI).

Table 7
Comparison of proposed method with six states arts models: Xception, NASNetMobile, InceptionResNetV2, MobileNetV2, ResNet50, and VGG16.

TL models	KA_{bc}	KP_{bc}	KR_{bc}	KF_{bc}	KM_{bc}	KK_{bc}
Xception	84.1 %	88.8 %	86.4 %	87.2 %	80.9 %	80.6 %
NASNetMobile	89.8 %	89.1 %	87.2 %	87.8 %	81.3 %	81.0 %
InceptionResNetV2	94.5 %	88.6 %	88.7 %	88.4 %	81.7 %	81.4 %
MobileNetV2	86.0 %	88.7 %	86.1 %	86.7 %	79.7 %	78.9 %
ResNet50	79.6 %	81.3 %	68.1 %	70.1 %	60.1 %	55.6 %
VGG16	87.2 %	88.8 %	87.1 %	87.8 %	81.4 %	81.2 %
VGG16-LSTM with SMOTETomek-LR-L	96.1 %	96.1 %	96.3 %	96.0 %	94.2 %	94.1 %

87.2 %, 87.8 %, 81.3 %, and 81.0 %, respectively. However, the proposed method (VGG16-LSTM with SMOTETomek-LR-L) outperformed all, achieving the highest 10-fold average KA_{bc}, KR_{bc}, KR_{bc}, KF_{bc}, KM_{bc}, and KK_{bc} of 96.1 %, 96.1 %, 96.3 %, 96.0 %, 94.2 %, and 94.1 %, respectively. This indicates that the proposed method outperformed all six SOTA models.

5. Discussion

BC is women's most often detected malignancy, posing a substantial global danger to women's health. BC has deep-reaching consequences that transcend beyond medical care, impacting families, communities, and people. However, early identification of BC may give rise to effective treatment and better results. ML/DL has emerged as a crucial tool in the screening and diagnosing of BC. Its promise resides not only in its capacity to analyze medical pictures but also in its capacity to relieve radiologists' workload and compensate for the difficulties that newcomers to the profession may experience. This union of technology and healthcare has enormous promise. Understanding medical pictures, such as mammograms, depends significantly on radiologists' skills. However, ML/DL models can improve this method by detecting small information in images that the human eye may miss. These models can see patterns and abnormalities that may suggest a possibility of BC, but they also give qualitative evaluations on their own. This change from qualitative to quantitative analysis has the potential to produce more precise and uniform outcomes. In terms of precision and effectiveness, these models have the potential to beat traditional techniques. They can filter through

massive volumes of data, detecting minute variations in tissue patterns and traits that may suggest BC. This skill may aid in the earlier detection of lesions, increasing the likelihood of effective treatment. Including ML and DL models in BC diagnosis does not seek to replace radiologists' knowledge but to supplement and augment it. These models are helpful tools that assist medical practitioners in their diagnostic attempts, allowing them to make better-informed judgments more effectively. Furthermore, ML/DL can bridge the gap between expert radiologists and those still learning about the field.

Various studies have been published to identify BC from USIs using ML or DL methods. References [18–20,23,25–32] have published work on two classes with ML or DL methods and achieved the highest accuracy of 98.84 % and AUC of 0.99 by Atrey et al. [28]. References [16,17,21,22,24] have published work on three classes (normal, malignant, and benign) using ML or DL methods and achieved the highest accuracy of 97.18 % and F1 score of 96.79 %. These studies successfully predict the BC from USIs. Most of the studies [16,18,20–23,25,26,29] have not applied the cross-validation method, which is an essential step in checking the robustness of the proposed system.

In the quest to detect BC utilizing contrast-enhanced ultrasonography (CES) recordings, Chen et al. [55] introduced a 3D CNN diagnostic model featuring domain-knowledge-guided temporal and channel attention modules. Their approach involved scrutinizing specific time intervals and discerning disparities between CEUS frames and corresponding USIs, unveiling distinctive patterns. Leveraging mask images, they achieved a sensitivity of 97.2 %, accuracy of 86.3 %, and an F1 score of 88.7 %, with an upper confidence interval (UCI) of 97.2 % and a lower confidence interval (LCI) of 72.5 %. In our proposed methodology, we trained the model using both original and mask images, yielding outstanding results: F1 score of 99.0 %, MCC of 98.9 %, kappa score of 98.9 %, and AUC of 100 %. Additionally, the 10-fold average F1 score reached 96.0 %, with MCC and kappa scores of 94.2 % and 94.1 %, respectively. This shows that our proposed method outperforms Chen et al.'s method. The proposed method achieves an LCI, HCI, and mean of 96.50 %, 99.75 %, and 98.13 %, respectively, better than Chen et al.'s method. A vital advantage of the proposed method is its ability to work with both USIs and mask images. This makes the proposed method more versatile and applicable to various tasks.

This work developed a TL-based method with LSTM to identify normal, malignant, and benign USIs, which helps to detect BC. We have also considered the dataset imbalance problem and applied the SMOTETomek resampling method to balance the dataset, which helps to improve the performance. ML classifiers were applied to classify the BC. To evaluate the performance of the imbalanced dataset, the F1 score, MCC, and Kappa Coefficient metrics are beneficial along with the K-fold method. The proposed method with VGG16-LSTM-SMOTETomek achieved the highest F1 score of 99 %, MCC and Kappa Coefficient of 98.9 % with an AUC of 1. Moreover, we have applied Grad-CAM for visualization, LIME for interpretability, and NAI and bootstrapping for calculating 95 % CIs with mean. The proposed method achieved LCI, HCI, and mean CI of 96.50 %, 99.75 %, and 98.13 %, respectively, with the NAI, while 95 % LCI of 93.81 %, an HCI of 96.00 %, and a bootstrap mean of 94.90 % with the bootstrap method.

From the above findings, the proposed method detects BC efficiently by integrating TL models with LSTM, SMOTETomek, and standard ML classifiers for detecting BC from USIs. Combining these strategies constitutes a holistic strategy for addressing issues associated with medical image analysis, using the strengths of many methodologies. The fusion of TL models with LSTM, SMOTETomek, and ML for BC detection from USIs surpassed state-of-the-art approaches in this study. First, pretrained TL pulls high-level information from USIs. These features capture the underlying patterns in USIs, which are crucial for adequately identifying BC. On the other hand, LSTM networks are particularly adept at dealing with sequential data, such as image sequences. Because USIs are recorded as pixel value sequences, LSTM can assist in identifying temporal correlations and patterns [55]. Combining these two approaches allows for a more in-depth examination of USIs, resulting in improved detection accuracy and a more robust and discriminative ultrasonic image representation. The regularization properties of LSTM aid in the prevention of overfitting during training. The model can distinguish between benign and malignant BC using this enhanced feature representation. SMOTETomek, on the other hand, solves the issue of imbalanced data classification by generating new minority class samples using SMOTE and then removing Tomek links, which are pairings of minority class instances that are too close in the feature space to be related to one another. This improves data distribution from minority classes, reduces overfitting, and improves classification model generalization.

The findings revealed that our proposed approach may improve radiologists' sensitivity and precision. The suggested technique could enhance the identification rate of early BC even more. The proposed work aimed to create a system that could detect minor alterations that radiologists might misread in an early BC stage. As a result, the use of DL in BC testing and diagnosis is crucial. Experts can use our suggested approach to evaluate breast lesions. The proposed method might aid in detecting early cancer and reduce the likelihood of misdiagnosis. Despite its promising performance, the suggested fusion approach has certain drawbacks. The effectiveness of this procedure is highly dependent on the quality and quantity of ultrasonic images. Scarcity of data or noise can limit the model's ability to train and generalise effectively. While DL models are accurate, their interior workings can be complex. This can make it challenging to grasp the model's decision-making process and discover biases. The model's efficacy may vary based on the USIs-capturing process, equipment type, and patient group.

6. Conclusion

In this study, the authors combined TL, LSTM networks, SMOTETomek, and ML classifiers for BC diagnosis using USIs, making a convincing argument for the combined power of many approaches. This study also revealed the utility of TL models as feature extractors for collecting subtle patterns and textures in USIs. This capacity to extract key features from complicated pictures improves later application of LSTM networks by permitting the modelling of temporal relationships, essential to comprehending the dynamic evolution of breast tissue features linked with cancerous progression. The SMOTETomek method also plays a vital role in correcting class imbalance, a key challenge in the medical domain.

The authors used an imbalanced USIs dataset and three TL models (MobileNetV2, ResNet50, and VGG16) with LSTM to extract features from the USIs, and these features were applied to the classifiers. The outcome of these methods was acceptable. The SMOTETomek method was applied to balance the dataset and achieved accuracy, F1 score, MCC, kappa score, and AUC of 99.0 %, 99.0 %, 98.9 %, 98.9 %, and 100.0 %, respectively, with VGG16-LSTM-SMOTETomek-LR-L model. The 10-fold method was applied and found that VGG16-LSTM-SMOTETomek-LR-L achieved 10-fold accuracy, F1 score, MCC, and kappa score of 96.1 %, 96.0 %, 94.2 %, and 94.1 %, respectively. Moreover, the Grad-CAM for visualization and the LIME method was applied for interpretability, which promotes confidence, model validation, and refinement. NAI and bootstrapping methods were also applied for calculating CIs. The VGG16-LSTM-SMOTETomek-LR-N classifier's lower, higher, and mean CIs were 96.50 %, 99.75 %, and 98.13 %, respectively, with the NAI method. On the other hand, using the bootstrap method, the VGG16-LSTM-SMOTETomek-SVCR model achieved a 95 % LCI of 93.81 %, an HCI of 96.00 %, and a bootstrap mean of 94.90 %.

In the future, the authors will integrate this method with user-friendly interfaces and decision assistance features, which are more appealing to medical practitioners. The authors will also test this method on different US datasets to detect breast cancer.

Authors contribution statement

Madhusduan G Lanjewar: Conceptualization, Methodology, Software, Validation, Investigation, Writing - Original Draft. Lalchand B Patle: Introduction, literature survey, corrections, and modifications to the original manuscript, and Kamini G Panchbhai: Introduction, Investigation, Performance analysis, and modifications of the original manuscript.

Ethical and informed consent for data used

Not applicable.

Data availability and access

The dataset available on following link: https://www.kaggle.com/datasets/aryashah2k/breast-ultrasound-images-dataset.

The Python code will be share on reasonable request.

CRediT authorship contribution statement

Madhusudan G. Lanjewar: Writing – review & editing, Writing – original draft, Visualization, Supervision, Software, Methodology, Conceptualization. **Kamini G. Panchbhai:** Writing – original draft, Resources, Investigation, Data curation. **Lalchand B. Patle:** Writing – original draft, Supervision, Investigation, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] X. Du, Y. Li, L. Fu, H. Chen, X. Zhang, Y. Shui, A. Zhang, X. Feng, M.R. Fu, Strategies in activating lymphatic system to promote lymph flow on lymphedema symptoms in breast cancer survivors: a randomized controlled trial, Front. Oncol. 12 (2022) 1015387, https://doi.org/10.3389/fonc.2022.1015387.

[2] E.M. Hussein Saeed, H.A. Saleh, E.A. Khalel, Classification of mammograms based on features extraction techniques using support vector machine, Comput. Sci. Inf. Technol. 2 (2020) 121–131, https://doi.org/10.11591/csit.v2i3.p121-131.

[3] M.S. Mahmoodi, S.A. Mahmoodi, Hybrid harmony search and genetic for fuzzy classification systems, J. Math. Comput. Sci. 10 (2014) 203–211, https://doi.org/10.22436/jmcs.010.03.06.

[4] K. Sathishkumar, M. Chaturvedi, P. Das, S. Stephen, P. Mathur, Cancer incidence estimates for 2022 & projection for 2025: result from national cancer registry programme, India, Indian J. Med. Res. 0 (2023), https://doi.org/10.4103/ijmr.ijmr_1821_22.

[5] BRCA Gene Mutations: Cancer Risk and Genetic Testing Fact Sheet - NCI, 2020. https://www.cancer.gov/about-cancer/causes-prevention/genetics/brca-fact-sheet. (Accessed 3 December 2023).

[6] menopause Menarche, And breast cancer risk: individual participant meta-analysis, including 118 964 women with breast cancer from 117 epidemiological studies, Lancet Oncol. 13 (2012) 1141–1151, https://doi.org/10.1016/S1470-2045(12)70425-4.

[7] CdcbreastCancer, What Are the Risk Factors for Breast Cancer? Centers for Disease Control and Prevention, 2023. https://www.cdc.gov/cancer/breast/basic_info/risk_factors.htm. (Accessed 3 December 2023).

[8] E.M. John, A.I. Phipps, A. Davis, J. Koo, Migration history, acculturation, and breast cancer risk in hispanic women, cancer epidemiology, Biomarkers & Prevention 14 (2005) 2905–2913, https://doi.org/10.1158/1055-9965.EPI-05-0483.

[9] Mammograms - Nci. https://www.cancer.gov/types/breast/mammograms-fact-sheet, 2023. (Accessed 3 December 2023).

[10] J. Dheeba, N. Albert Singh, S. Tamil Selvi, Computer-aided detection of breast cancer on mammograms: a swarm intelligence optimized wavelet neural network approach, J. Biomed. Inf. 49 (2014) 45–52, https://doi.org/10.1016/j.jbi.2014.01.010.

[11] L. Provencher, J.C. Hogue, C. Desbiens, B. Poirier, E. Poirier, D. Boudreau, M. Joyal, C. Diorio, N. Duchesne, J. Chiquette, Is clinical breast examination important for breast cancer detection? Curr. Oncol. 23 (2016) 332–339, https://doi.org/10.3747/co.23.2881.

[12] Breast Magnetic Resonance Imaging (MRI), 2021. https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/breast-mri. (Accessed 3 December 2023).

[13] Y. Zhang, L. Deng, H. Zhu, W. Wang, Z. Ren, Q. Zhou, S. Lu, S. Sun, Z. Zhu, J. M. Gorriz, S. Wang, Deep learning in food category recognition, Inf. Fusion 98 (2023) 101859, https://doi.org/10.1016/j.inffus.2023.101859.

[14] R. Wang, Comparison of decision tree, random forest and linear discriminant analysis models in breast cancer prediction, J. Phys.: Conf. Ser. 2386 (2022) 012043, https://doi.org/10.1088/1742-6596/2386/1/012043.

[15] S. Lu, Z. Zhu, J.M. Gorriz, S. Wang, Y. Zhang, NAGNN: classification of COVID-19 based on neighboring aware representation from deep graph neural network, Int J of Intelligent Sys 37 (2022) 1572–1598, https://doi.org/10.1002/int.22686.

[16] B. Hajipour Khire Masjidi, S. Bahmani, F. Sharifi, M. Peivandi, M. Khosravani, A. Hussein Mohammed, CT-ML: diagnosis of breast cancer based on ultrasound images and time-dependent feature extraction methods using contourlet transformation and machine learning, Comput. Intell. Neurosci. (2022), https://doi.org/10.1155/2022/1493847, 2022) 1–15.

[17] H. Liu, G. Cui, Y. Luo, Y. Guo, L. Zhao, Y. Wang, A. Subasi, S. Dogan, T. Tuncer, Artificial intelligence-based breast cancer diagnosis using ultrasound images and grid-based deep feature generator, Int. J. Graph Multimed. 15 (2022) 2271–2282, https://doi.org/10.2147/IJGM.S347491.

[18] J. Baek, A.M. O'Connell, K.J. Parker, Improving breast cancer diagnosis by incorporating raw ultrasound parameters into machine learning, Mach. Learn.: Sci. Technol. 3 (2022) 045013, https://doi.org/10.1088/2632-2153/ac9bcc.

[19] A. Boulenger, Y. Luo, C. Zhang, C. Zhao, Y. Gao, M. Xiao, Q. Zhu, J. Tang, Deep learning-based system for automatic prediction of triple-negative breast cancer from ultrasound images, Med. Biol. Eng. Comput. 61 (2023) 567–578, https://doi.org/10.1007/s11517-022-02728-4.

[20] Y. Gu, W. Xu, B. Lin, X. An, J. Tian, H. Ran, W. Ren, C. Chang, J. Yuan, C. Kang, Y. Deng, H. Wang, B. Luo, S. Guo, Q. Zhou, E. Xue, W. Zhan, Q. Zhou, J. Li, P. Zhou, M. Chen, Y. Gu, W. Chen, Y. Zhang, J. Li, L. Cong, L. Zhu, H. Wang, Y. Jiang, Deep learning based on ultrasound images assists breast lesion diagnosis in China: a multicenter diagnostic study, Insights Imaging 13 (2022) 124, https://doi.org/10.1186/s13244-022-01259-8.

[21] N. Sirjani, M. Ghelich Oghli, M. Kazem Tarzamni, M. Gity, A. Shabanzadeh, P. Ghaderi, I. Shiri, A. Akhavan, M. Faraji, M. Taghipour, A novel deep learning model for breast lesion classification using ultrasound Images: a multicenter data evaluation, Phys. Med. 107 (2023) 102560, https://doi.org/10.1016/j.ejmp.2023.102560.

[22] S.-H. Chen, Y.-L. Wu, C.-Y. Pan, L.-Y. Lian, Q.-C. Su, Breast ultrasound image classification and physiological assessment based on GoogLeNet, Journal of Radiation Research and Applied Sciences 16 (2023) 100628, https://doi.org/10.1016/j.jrras.2023.100628.

[23] W.-J. Shen, H.-X. Zhou, Y. He, W. Xing, Predicting female breast cancer by artificial intelligence: Combining clinical information and BI-RADS ultrasound descriptors, WFUMB Ultrasound Open 1 (2023) 100013, https://doi.org/10.1016/j.wfumbo.2023.100013.

[24] J. Liao, Y. Gui, Z. Li, Z. Deng, X. Han, H. Tian, L. Cai, X. Liu, C. Tang, J. Liu, Y. Wei, L. Hu, F. Niu, J. Liu, X. Yang, S. Li, X. Cui, X. Wu, Q. Chen, A. Wan, J. Jiang, Y. Zhang, X. Luo, P. Wang, Z. Cai, L. Chen, Artificial intelligence-assisted ultrasound image analysis to discriminate early breast cancer in Chinese population: a retrospective, multicentre, cohort study, eClinicalMedicine 60 (2023) 102001, https://doi.org/10.1016/j.eclinm.2023.102001.

[25] Q. Wang, H. Chen, G. Luo, B. Li, H. Shang, H. Shao, S. Sun, Z. Wang, K. Wang, W. Cheng, Performance of novel deep learning network with the incorporation of the automatic segmentation network for diagnosis of breast cancer in automated breast ultrasound, Eur. Radiol. 32 (2022) 7163–7172, https://doi.org/10.1007/s00330-022-08836-x.

[26] H. Taleghamar, S.A. Jalalifar, G.J. Czarnota, A. Sadeghi-Naini, Deep learning of quantitative ultrasound multi-parametric images at pre-treatment to predict breast cancer response to chemotherapy, Sci. Rep. 12 (2022) 2244, https://doi.org/10.1038/s41598-022-06100-2.

[27] E. Fleury, K. Marcomini, Performance of machine learning software to classify breast lesions using BI-RADS radiomic features on ultrasound images, Eur Radiol Exp 3 (2019) 34, https://doi.org/10.1186/s41747-019-0112-7.

[28] K. Atrey, B.K. Singh, N.K. Bodhey, Multimodal classification of breast cancer using feature level fusion of mammogram and ultrasound images in machine learning paradigm, Multimed. Tool. Appl. (2023), https://doi.org/10.1007/s11042-023-16414-6.

[29] T. Wu, L.R. Sultan, J. Tian, T.W. Cary, C.M. Sehgal, Machine learning for diagnostic ultrasound of triple-negative breast cancer, Breast Cancer Res. Treat. 173 (2019) 365–373, https://doi.org/10.1007/s10549-018-4984-7.

[30] H. Zhang, W. Cao, L. Liu, Z. Meng, N. Sun, Y. Meng, J. Fei, Noninvasive prediction of node-positive breast cancer response to presurgical neoadjuvant chemotherapy therapy based on machine learning of axillary lymph node ultrasound, J. Transl. Med. 21 (2023) 337, https://doi.org/10.1186/s12967-023-04201-8.

[31] R. Karthiga, K. Narasimhan, Automated diagnosis of breast cancer from ultrasound images using diverse ML techniques, Multimed. Tool. Appl. 81 (2022) 30169–30193, https://doi.org/10.1007/s11042-022-12933-w.

[32] V. Romeo, R. Cuocolo, R. Apolito, A. Stanzione, A. Ventimiglia, A. Vitale, F. Verde, A. Accurso, M. Amitrano, L. Insabato, A. Gencarelli, R. Buonocore, M.R. Argenzio, A.M. Cascone, M. Imbriaco, S. Maurea, A. Brunetti, Clinical value of radiomics and machine learning in breast ultrasound: a multicenter study for differential diagnosis of benign and malignant lesions, Eur. Radiol. 31 (2021) 9511–9519, https://doi.org/10.1007/s00330-021-08009-2.

[33] S. Lu, S.-H. Wang, Y.-D. Zhang, Detection of abnormal brain in MRI via improved AlexNet and ELM optimized by chaotic bat algorithm, Neural Comput. Appl. 33 (2021) 10799–10811, https://doi.org/10.1007/s00521-020-05082-4.
[34] Breast Ultrasound Images Dataset, (n.d.). https://www.kaggle.com/datasets/aryashah2k/breast-ultrasound-images-dataset (accessed December 3, 2023).
[35] W. Al-Dhabyani, M. Gomaa, H. Khaled, A. Fahmy, Dataset of breast ultrasound images, Data Brief 28 (2020) 104863, https://doi.org/10.1016/j.dib.2019.104863.
[36] M.G. Lanjewar, K.G. Panchbhai, Convolutional neural network based tea leaf disease prediction system on smart phone using paas cloud, Neural Comput. Appl. 35 (2023) 2755–2771, https://doi.org/10.1007/s00521-022-07743-y.
[37] F. Chollet, Xception: Deep Learning with Depthwise Separable Convolutions, 2016, https://doi.org/10.48550/ARXIV.1610.02357.
[38] C. Szegedy, S. Ioffe, V. Vanhoucke, A. Alemi, Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning, 2016, https://doi.org/10.48550/ARXIV.1602.07261.
[39] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, MobileNets, Efficient Convolutional Neural Networks for Mobile Vision Applications, 2017, https://doi.org/10.48550/ARXIV.1704.04861.
[40] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, MobileNetV2: inverted residuals and linear bottlenecks, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, Salt Lake City, UT, 2018, pp. 4510–4520, https://doi.org/10.1109/CVPR.2018.00474.
[41] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, 2015, https://doi.org/10.48550/ARXIV.1512.03385.
[42] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, 2014, https://doi.org/10.48550/ARXIV.1409.1556.
[43] X. Wang, T. Huang, K. Zhu, X. Zhao, LSTM-based broad learning system for remaining useful life prediction, Mathematics 10 (2022) 2066, https://doi.org/10.3390/math10122066.
[44] S. Saxena, What Is LSTM? Introduction to Long Short-Term Memory, Analytics Vidhya, 2021. https://www.analyticsvidhya.com/blog/2021/03/introduction-to-long-short-term-memory-lstm/. (Accessed 3 December 2023).
[45] M. Phi, Illustrated Guide to LSTM's and GRU's: A Step by Step Explanation, Medium, 2020. https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21. (Accessed 3 December 2023).
[46] D. Thakur, LSTM and its Equations, Medium, 2018. https://medium.com/@divyanshu132/lstm-and-its-equations-5ee9246d04af. (Accessed 3 December 2023).
[47] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, Jair 16 (2002) 321–357, https://doi.org/10.1613/jair.953.
[48] A. Fernandez, S. Garcia, F. Herrera, N.V. Chawla, SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary, Jair 61 (2018) 863–905, https://doi.org/10.1613/jair.1.11192.
[49] R.M. AlZoman, M.J.F. Alenazi, A comparative study of traffic classification techniques for smart city networks, Sensors 21 (2021) 4677, https://doi.org/10.3390/s21144677.
[50] G. Rau, Y.-S. Shih, Evaluation of Cohen's kappa and other measures of inter-rater agreement for genre analysis and other nominal data, J. Engl. Acad. Purp. 53 (2021) 101026, https://doi.org/10.1016/j.jeap.2021.101026.
[51] M.G. Lanjewar, K.G. Panchbhai, P. Charanarur, Lung cancer detection from CT scans using modified DenseNet with feature selection methods and ML classifiers, Expert Syst. Appl. 224 (2023) 119961, https://doi.org/10.1016/j.eswa.2023.119961.
[52] J.R. Landis, G.G. Koch, The measurement of observer agreement for categorical data, Biometrics 33 (1977) 159, https://doi.org/10.2307/2529310.
[53] X. Bouthillier, P. Delaunay, M. Bronzi, A. Trofimov, B. Nichyporuk, J. Szeto, N. Sepah, E. Raff, K. Madan, V. Voleti, S.E. Kahou, V. Michalski, D. Serdyuk, T. Arbel, C. Pal, G. Varoquaux, P. Vincent, Accounting for Variance in Machine Learning Benchmarks, 2021, https://doi.org/10.48550/ARXIV.2103.03098.
[54] W. Panichkitkosolkul, Confidence intervals for the coefficient of variation in a normal distribution with a known population mean, Journal of Probability and Statistics 2013 (2013) 1–11, https://doi.org/10.1155/2013/324940.
[55] C. Chen, Y. Wang, J. Niu, X. Liu, Q. Li, X. Gong, Domain knowledge powered deep learning for breast cancer diagnosis based on contrast-enhanced ultrasound videos, IEEE Trans. Med. Imag. 40 (2021) 2439–2451, https://doi.org/10.1109/TMI.2021.3078370.
[56] M.G. Lanjewar, O.L. Gurav, Convolutional Neural Networks based classifications of soil images, Multimed. Tool. Appl. 81 (2022) 10313–10336, https://doi.org/10.1007/s11042-022-12200-y.
[57] M.G. Lanjewar, A.Y. Shaikh, J. Parab, Cloud-based COVID-19 disease prediction system from X-Ray images using convolutional neural network on smartphone, Multimed. Tool. Appl. 82 (2023) 29883–29912, https://doi.org/10.1007/s11042-022-14232-w.