Revista de Informática Teórica e Aplicada - RITA - ISSN 2175-2745
Vol. 32, Num. 1 (2025) 11-17

RESEARCH ARTICLE

Transfer Learning and Handcrafted Features Ensembles for Ultrasound Breast Cancer Image Classification

Comitês de Aprendizado por Transferência e Características Manuais para Classificação de Câncer de Mama em Imagens de Ultrassom

Vanessa Kaplum Foleis[1]*, Bianca Aparecida Andrade[1], Henrique Bergamo Shigihara[1], Dimas Augusto Mendes Lemes[2], José Guilherme Picolo[1], Bernardo Feijó Junqueira[1], Guilherme Ribeiro Sales[1], Valentino Corso[1], Cides S. Bezerra[1]

Abstract: Breast cancer is the most commonly diagnosed cancer in women. Its diagnosis via ultrasound imaging largely depends on the technical skill of the radiologist. This study developed a binary classification system for breast lesions, combining transfer learning models and handcrafted features in ultrasound images. Pre-trained CNNs like InceptionV3, EfficientNetB4, ResNet50, and VGG16 were used, along with SVM-classified handcrafted features. Models were individually analyzed and combined using late-fusion ensembles. ResNet50 achieved an F1-score of 81.97%. The best late-fusion ensemble model reached an F1-score of 83.90%. In the cross-dataset evaluation, the top late-fusion ensemble model in the development dataset scored an F1-score of 88.70% and 78.20% in the test BUSI and BUID datasets, respectively. These results emphasize the robust potential of a late-fusion ensemble that combines CNN transfer learning and handcrafted features to classify breast lesions in ultrasound images.
Keywords: Breast Cancer — Ultrasound Images — Classification — Transfer Learning — Late-Fusion Ensemble — Cross-dataset Analysis

Resumo: O câncer de mama é o câncer mais comumente diagnosticado em mulheres. Seu diagnóstico por meio de imagens de ultrassom depende, em grande parte, da habilidade técnica do radiologista. Este estudo desenvolveu um sistema de classificação binária para lesões mamárias, combinando modelos de aprendizado por transferência e características manuais em imagens de ultrassom. Redes Neurais Convolucionais (CNNs) pré-treinadas, como InceptionV3, EfficientNetB4, ResNet50 e VGG16, foram utilizadas, juntamente com características manuais classificadas por SVM. Os modelos foram analisados individualmente e combinados usando comitês de fusão tardia. A ResNet50 alcançou um F1-score de 81,97%. O melhor comitê de fusão tardia alcançou um F1-score de 83,90%. Na avaliação cruzada de bases de dados, o melhor comitê obteve F1-score de 88,70% e 78,20% nas bases de dados BUSI e BUID, respectivamente. Os resultados enfatizam o robusto potencial do comitê de fusão tardia que combina aprendizado por transferência e características manuais para classificar lesões mamárias em imagens de ultrassom.
Palavras-Chave: Câncer de Mama — Ultrassom — Classificação — Apredizado por Transferência — Comitê de Fusão Tardia — Avaliação Cruzada de Bases de Dados

[1] *CPQD - Centro de Pesquisa e Desenvolvimento, Brazil*
[2] *Pontifícia Universidade Católica de Campinas (PUCC), Brazil*
***Corresponding author**: vanessakaplum@gmail.com
DOI: http://dx.doi.org/10.22456/2175-2745.143357 • **Received:** 23/10/2024 • **Accepted:** 17/12/2024

1. Introduction

Breast cancer is the most commonly diagnosed cancer in women and the leading cause of cancer-related deaths. Each year, it is estimated that 2.3 million new cases of breast cancer are diagnosed globally [1]. Diagnosis of breast cancer is usually performed using mammography, digital breast tomosynthesis (DBT), and ultrasound imaging [2].

Breast ultrasound stands out for its low cost, wide availability, and not using ionizing radiation [3]. Benign and malignant lesions can be differentiated using ultrasound elastography, which assesses tissue stiffness and vascular permeability

[4]. However, diagnosis based on ultrasound directly depends on the technical skill of the radiologist [3]. Other factors can result in misdiagnosis, such as distractions and fatigue [5].

Computer-aided diagnosis systems (CAD systems) assist radiologists in decision-making, especially during lesion classification, contributing to improved and definitive decision. By analyzing image patterns, CAD systems enhance diagnosis accuracy, leading to improved patient outcomes through early disease detection [6].

Multiple deep learning (DL) methods were employed to classify breast lesions present in ultrasound images [7]. Convolutional neural networks (CNNs) stand out as one of the most popular methods, along with DL ensembles [8]. Classifiers can be further improved by incorporating handcrafted fearures, such as texture and shape attributes [9]. However, it is crucial to evaluate their generalization capability on other datasets with varying levels of noise and signal variations to determine their practical utility in real-world breast ultrasound imaging.

This research aimed to develop a binary classification system for breast lesions by combining models utilizing transfer learning and handcrafted features in the BUS-BRA dataset. The proposed system comprises CNN pre-trained models classified by multilayer perceptrons (MLPs), including InceptionV3, EfficientNetB4, ResNet50, and VGG16. Handcrafted features classified by support vector machine (SVM) were also adopted, specifically general-purpose texture descriptors and radiomic features. Each model was individually analyzed and then combined through late-fusion ensembles using the sum rule. The top ensemble models were chosen for cross-dataset evaluation on the BUSI and BUID datasets[10],[11]. Performance was evaluated using various metrics.

The major contributions of this work are as follows:

i) Novel combinations of deep learning and handcrafted features are proposed to assess improvements in binary classification.

ii) The robustness and generalization of late-fusion ensemble models are tested through cross-dataset evaluation on independent datasets.

This work is structured as follows: Section 2 presents a concise review. Section 3 covers datasets, image preprocessing, transfer learning methods, SVM classification, late-fusion ensembles, and cross-dataset evaluation. Section 4 presents the results and discussion. Finally, Section 5 concludes the work and suggests directions for future research.

2. Related Work

An overview of prior research in this domain is provided.

Among these studies, Ali et al. [12] proposed a breast cancer classification model that employs multiple CNNs (InceptionV3, ResNet50, and DenseNet121) with the BUSI dataset. The F1-scores for benign and malignant images were 90% and 89%, respectively. The study demonstrated that their meta-learning ensemble approach significantly improved accuracy, particularly enhancing the detection of tumors.

In a similar effort, Moon et al. [13] introduced a CAD system employing image fusion and ensemble CNN architectures (VGGNet, ResNet, DenseNet) for breast ultrasound tumor diagnosis. The method achieved an F1-score of 89.36% on a private dataset and 91.14% on the BUSI dataset. Their findings demonstrated that integrating multiple features yields superior results compared to using individual features alone.

Another study by Cruz-Ramos et al. [9] proposed a hybrid CAD system that combines CNN (DenseNet201) with handcrafted features (Histogram of Oriented Gradients (HOG)-based, LBP, perimeter area, area, eccentricity, and circularity). Applied to the BUSI dataset, their approach achieved an F1-score of 96%. The study highlighted the importance of combining deep learning and handcrafted features extracted from regions of interest (ROI).

Abhisheka et al. [14] also combined CNN-extracted features (ResNet50) with handcrafted features (HOG). Using an SVM classifier, their method achieved an F1-score of 84.28% on the BUSI dataset. Their study demonstrated that feature integration outperforms individual features.

Continuing this line of investigation, Abhisheka et al. [15] combined ResNet50 with handcrafted features, including HOG, Gray Level Co-occurrence Matrix (GLCM), and Local Binary Pattern (LBP). Their method achieved an F1-score of 87.18% on the BUSI dataset, showing that feature integration enhances model performance.

Nanni et al. [16] created an ensemble model, named TakhisisNet, which combines various modified CNNs with the original using the sum rule. Additionally, the model's performance was further enhanced by integrating it with some handcrafted features. It showed promising results, obtaining an AUC of 0.877, while GoogleNet achieved an AUC of 0.822.

Mo et al. [17] combined vision transformers with CNNs in the HoVer-Trans model, which was evaluated across three distinct datasets. On the BUSI dataset, this method achieved an F1-score of 87.20%. The model eliminated the need for a predefined ROI and outperformed several state-of-the-art methods.

3. Methods

3.1 Software Details

The code was implemented in Python 3.11. The models were implemented with the following Python tools: NumPy, Pandas, Matplotlib, Pillow, OpenCV, Scikit-Learn, Scikit-Image, Imblearn, TensorFlow, and Keras.

3.2 Datasets

The BUS-BRA is a public dataset comprising 1875 breast ultrasound images collected by a Brazilian National Cancer Institute in Rio de Janeiro, Brazil, including 1268 benign lesion images and 607 malignant ones [18]. The BUSI dataset, publicly available from the Baheya Hospital in Cairo, Egypt, comprises 780 breast ultrasound images, including 487 images of benign lesions, 210 images of malignant lesions, and 133 images without lesions [10]. Normal images are excluded.

The BUID dataset is a public database of breast ultrasound images collected by Shahid Beheshti University of Medical Science in Tehran, Iran, including 109 benign and 123 malignant lesion images, respectively [11].

3.3 Images Preprocessing

The grayscale images in the BUS-BRA, BUSI, and BUID datasets have different dimensions. Therefore, resizing them to fit the input layer size of the models is essential. Thus, the grayscale images were preprocessed in three distinct methods: Input A, Input B, and Input C, according to Fig. 1. Input A: The original grayscale image was directly resized to 224 x 224 pixels. Input B: The ROI of the original grayscale image was cropped based in the binary mask, then was directly resized to 224 x 224 pixels. Input C: The ROI crops were resized proportionally to the longest side, reaching a length of 224 pixels, with circular padding added on both sides of the shorter dimension to achieve a length of 224 pixels. Then, the image was converted to a 3-channel RGB format [13].

3.4 Transfer learning based classification

Convolutional neural networks (CNNs) are commonly used in image classification tasks, leveraging feature detection for improved generalization. Due to their numerous parameters, ample labeled data is necessary during training to mitigate overfitting. Transfer learning utilizes pre-trained models as feature extractors, beneficial for applications with limited datasets. Most CNN architectures comprise feature extraction layers and an MLP classifier, with features from the last extraction layer often used for classification.

The rationale behind transfer learning is that considering the model was pre-trained with a large dataset to predict among many classes, the features are expected to be generic enough to classification new, previously unseen classes. InceptionV3, EfficientNetB4, ResNet50, and VGG16 are commonly used backbones for image classification.

In this work, we propose evaluating different backbones in an end-to-end neural network for breast tumor classification. The backbones evaluated were ResNet50, EfficientNetB4, InceptionV3, and VGG16, pre-trained on the ImageNet dataset.

The backbone is used as the feature extractor, followed by a custom MLP classifier. The architecture of the custom MLP classifier was optimized by adjusting parameters such as the number of layers, the number of neurons per layer, and the dropout rate using Keras Tuner for hyperparameter optimization. For the EfficientNetB4 and InceptionV3, the output layers were set as MLP1 (a dense layer with 512 neurons, ReLu activation function, and L2 regularization, and a dropout layer with 40% dropout rate) and MLP2 (a dense layer with 1024 neurons, ReLu activation function, and L2 regularization), respectively, while MLP3 (a dense layer with 1024 neurons, ReLu activation function, and L2 regularization, and a dropout layer with 30% dropout rate) was employed for ResNet50 and VGG16.

The model was trained for 100 epochs at most. Early stopping was used to prevent overfitting, halting training after 10 epochs without improvement in a held-out validation set. 90% of the training data was used for model fitting, while 10% was used for validation.

During training, the weights in the backbone were kept frozen. The MLP classifier parameters were updated by backpropagation, minimizing the binary cross-entropy loss function with the ADAM optimizer. The mini-batch size was 32 images. Learning rate decay was employed to prevent getting stuck in local minima, starting at 10e-3 and reducing by 90% after five epochs without significant loss function improvement (10e-7).

Data augmentation was used to increase variability and prevent overfitting. The following transformations were randomly applied during training, generating different variations of the input images each epoch: rotations (up to 20 degrees), shifts (up to 20%), shear changes, zooms (80-120%), and flips.

3.5 SVM Classification of Handcrafted Features

Deep learning backbones extract valuable features but can be computationally expensive, with memory and processing demands proportional to their size and parameters. In contexts where these requirements are impractical, handcrafted features can be used for image classification.

Several radiomic features have been proposed for radiological image classification, detailing characteristics of regions of interest like tumors and clinically significant lesions. These features encompass texture, size, shape, and heterogeneity measurements. Commonly utilized features include first-order statistics of pixel data (FO), gray level run length matrix (GLRLM), gray level size zone matrix (GLSZM), gray level dependence matrix (GLDM), neighboring gray-tone difference matrix (NGTDM), and lesion shape measurements (SHAPE2D) [19].

Radiologists often evaluate the texture of a lesion in an image to assess its potential malignancy. In the texture classification literature, some well-known general-purpose texture descriptors have been successfully used in various classification applications [20].

In this work, we propose evaluating radiomic features and general-purpose texture descriptors in breast tumor classification. The following radiomic features were used: FO, GLRLM, GLSZM, GLDM, NGTDM, and SHAPE2D. The radiomic features were extracted with the PyRadiomics library, with default parameter values.

The general-purpose texture descriptors we evaluated were LBP (P=8, R=2, method="nri_uniform"), local phase quantization (LPQ; winSize=3, STFT window, decorrelation transform), and GLCM (distances=[1, 2, 3, 4], angles=[0, 45, 90, 135, 180, 225, 270, 315], properties: contrast, dissimilarity, homogeneity, energy, and correlation). Both LBP and GLCM were extracted with the Scikit-image library [21, 22]. The LPQ descriptor was implemented in Python [23].

The SVM with the RBF kernel was used to classify the image descriptors [24]. Its hyperparameters were tuned via

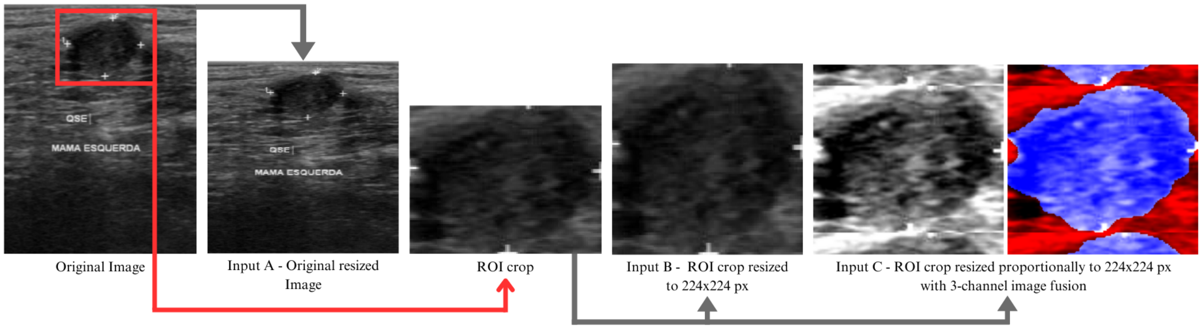

Figure 1. Summary of the preprocessing methods evaluated.

grid-search (C=[1, 10, 100, 1000], gamma=[2e-3, 2e-2, 2e-1, 1/(n_features * var(x)), 1/n_features]). Considering that the SVM requires the input data to be normalized in the same scale, the features were normalized accordingly. LBP and LPQ were normalized with z-score. All the other features were normalized with a quantile-based transform (500 quantiles), so they follow a uniform distribution.

The BUS-BRA dataset is not balanced, which can negatively impact classification performance. There are 1268 benign and 607 malignant lesions. The Synthetic Minority Over-sampling (SMOTE) technique was used to rebalance the dataset. It creates synthetic samples for the minority class derived from a linear combination of a reference sample and up to k nearest neighbors from the same class (k=7).

3.6 Late-fusion Ensemble

Distinct features can improve classification performance due to their complementary nature. Late-fusion, a type of ensemble learning, is a common method employed for this purpose [25]. It involves training individual classifiers for different feature sets and combining their probability estimates using fusion rules like sum or product. For instance, with the sum rule, probabilities from multiple classifiers for each class are added, and the class with the highest summed probability determines the final decision. This study explores the complementarity between transfer learning features and handcrafted features in breast tumor classification. Classifier probabilities from previous sections (Sections D and E) were combined using the sum rule, and the final decision was determined by the class with the highest sum. All $2^{13}-1$ possible classifier combinations were evaluated.

3.7 Cross-dataset analysis

Traditional machine learning evaluation involves training and testing models on separate partitions of the same dataset. However, this approach has drawbacks, as some models may not perform well in real-world scenarios due to domain shifts in the same problem. Such shifts arise from variations in signal and noise patterns among datasets. A recent study recommends using separate datasets for training (development dataset) and testing (test dataset) to assess model robustness[26].

In this work, we have developed the models using the BUS-BRA dataset. The top-performing models in the BUS-BRA dataset were selected for cross-dataset evaluation. For this, the selected models were retrained with the full BUS-BRA dataset and then evaluated with the full BUSI and BUID datasets.

3.8 Model evaluation

The 3-fold cross-validation technique was employed to validate the models, allowing for a robust assessment of its generalization capability by dividing the dataset into three distinct parts. Within each fold, the training data was split into 90% for training and 10% for validation.

The model performance was evaluated by six different metrics, namely the accuracy (ACC), precision (PREC), sensitivity (SEN), F1-score (F1), specificity (SPEC), and area under the receiver ROC curve (AUC). The F1-score was chosen as the primary metric for model comparison.

4. Results and Discussion

The binary classification of breast lesion ultrasound images was performed using three distinct preprocessing methods for the input image (Input A, Input B, and Input C). The images were classified using pre-trained CNNs as feature extractors, specifically EfficientNetB4, InceptionV3, ResNet50, and VGG16. Classification results are shown in Table 1.

ResNet50 showed superior classification performance among the architectures across all three preprocessing methods. As further evidenced in Table 1, modifications to the preprocessing methods increased the AUC values for ResNet50.

Another highlight was the progressive improvement in performance across all models as the preprocessing methods were refined. The F1-score for the original images was lower than the ROI-cropped images, indicating that noise removal enhanced the classification models evaluated.

Similarly, the ROI-cropped images that underwent proportional resizing demonstrated an increase in the F1-score metric, indicating that preserving the shape of the lesion is advantageous for the classification models evaluated. Thus, optimizing image preprocessing significantly influences the performance of classification models.

The results obtained are comparable to those published by [18], which also demonstrated that cropping and propor-

Table 1. Evaluation metrics for EfficientNetB4, InceptionV3, ResNet50 and VGG16 networks obtained from Input A, B and C.

Inputs	Networks	Evaluation metrics					
		ACC (%)	PREC (%)	SEN (%)	F1 (%)	SPEC (%)	AUC
Input A	EfficientNetB4	77.49 ± 0.86	71.44 ± 1.80	51.08 ± 5.85	59.31 ± 3.52	90.14 ± 1.74	0.83 ± 0.011
	InceptionV3	74.34 ± 0.86	62.91 ± 0.51	50.58 ± 6.30	55.86 ± 3.76	85.72 ± 1.77	0.76 ± 0.007
	ResNet50	77.06 ± 0.52	65.48 ± 1.16	61.77 ± 2.04	**63.54 ± 0.95**	84.38 ± 1.22	0.81 ± 0.008
	VGG16	74.98 ± 0.76	61.83 ± 1.83	59.80 ± 2.58	60.73 ± 1.05	82.25 ± 1.91	0.80 ± 0.014
Input B	EfficientNetB4	79.52 ± 1.17	72.32 ± 3.83	60.13 ± 2.84	65.52 ± 1.39	88.80 ± 2.61	0.85 ± 0.015
	InceptionV3	77.97 ± 1.39	68.91 ± 2.88	59.80 ± 1.33	62.96 ± 6.61	86.67 ± 4.74	0.84 ± 0.012
	ResNet50	82.56 ± 0.98	72.72 ± 3.12	73.30 ± 3.48	**73.38 ± 1.28**	86.51 ± 2.45	0.88 ± 0.005
	VGG16	81.06 ± 1.65	69.65 ± 2.16	73.47 ± 3.99	71.48 ± 2.91	84.70 ± 0.93	0.86 ± 0.009
Input C	EfficientNetB4	87.47 ± 0.30	81.70 ± 2.66	79.25 ± 3.10	80.36 ± 0.54	91.41 ± 1.74	0.92 ± 0.002
	InceptionV3	86.34 ± 1.05	79.71 ± 3.00	77.76 ± 0.35	78.68 ± 1.29	90.45 ± 1.72	0.92 ± 0.005
	ResNet50	88.64 ± 0.59	84.47 ± 2.98	79.74 ± 1.80	**81.97 ± 0.58**	92.90 ± 1.66	0.93 ± 0.006
	VGG16	85.33 ± 1.55	76.18 ± 3.32	79.90 ± 1.64	77.94 ± 1.82	87.93 ± 2.36	0.91 ± 0.009

tionally resizing the ROI in the BUS-BRA dataset yielded better classification results with ResNet50, achieving accuracy, sensitivity, specificity, and AUC values of 85.9%, 80.8%, 88.3%, and 0.92, respectively.

General-purpose texture descriptors, namely GLCM, LBP, and LPQ, achieved F1-scores of 63.70%, 40.30%, and 51.90%, respectively (Table 2). GLCM, which describes pixel spatial relationships with varying gray levels, had the highest F1-score among these descriptors.

Radiomic features, quantitative characteristics extracted from the delineated ROIs, are typically applied to breast imaging diagnosis. For the radiomic features, according to Table 2, the F1-score for FO, GLDM, GLRLM, GLSZM, NGTDM e SHAPE2D was 53.00%, 57.00%, 60.20%, 49.00%, 55.90%, and 80.50%, respectively.

The shape-based feature demonstrated superior performance. The SHAPE2D descriptor captures geometric attributes of ROIs, such as diameters, axes, compactness, and sphericity [19]. Spiculations and a taller-than-wide orientation are commonly observed in malignant lesions [3]. These characteristics captured by SHAPE2D may enhance model generalization performance.

The ensemble method combines multiple models to create a more accurate and comprehensive generalized model [25]. Thus, integrating models trained with handcrafted features and transfer learning features can contribute to more robust models for breast cancer classification [9].

A late-fusion ensemble was developed for the BUS-BRA dataset, combining probabilities from four different pre-trained CNN models, three general-purpose texture descriptors, and six radiomic descriptors. This resulted in 8,191 combinations.

The most effective combination consists of the GLCM general-purpose texture descriptor, GLDM, GLRLM, NGTDM, and SHAPE2D radiomic descriptors, and features extracted by InceptionV3, EfficientNetB4, ResNet50, and VGG16. For this combination, accuracy, precision, sensitivity, F1-score, and specificity were 89.90%, 85.70%, 82.20%, 83.90%, and 93.50%, respectively (Table 3).

The evaluation metrics for the top-performing late-fusion ensemble surpassed those achieved by each set of handcrafted and transfer learning features individually. Moreover, it demonstrated superior accuracy, sensitivity, and specificity compared to the results reported by [18].

The late-fusion ensemble was developed using the BUS-BRA dataset. A cross-dataset analysis was performed to evaluate the model's robustness, wherein the model was assessed with unseen data not used in development namely the BUSI and BUID datasets. For the optimal combination within the late-fusion ensemble with the BUS-BRA dataset, the F1-score of the BUSI and BUID datasets were 88.70% and 78.20%, respectively. Thus, the F1-score obtained for the BUSI and BUID datasets was close to that achieved during model development (Table 3).

The results suggest that the proposed late-fusion ensemble model exhibited robustness when evaluated on unseen data from both BUSI and BUID datasets. This is particularly significant because evaluating the model on a new dataset simulates novel real-world situations with different noise and signal patterns. Therefore, testing the model on another dataset is crucial to assess its generalization ability.

The literature emphasizes several promising techniques for diagnosing lesions in ultrasound images. Table 3 compares the proposed ensemble model with recent state-of-the-art methods. In cross-dataset evaluations with the BUSI dataset, our ensemble model outperformed other systems that employ both deep and handcrafted features [14], [15]. However, it was inferior to that reported by [9]. On the other hand, the performance of our ensemble model was comparable to that of HoVer-Trans, an advanced vision transformer model [17].

The approach shows promise but is limited by ROI annotation variability and applicability to complex diagnostic tasks. Addressing these issues could enhance its robustness.

Table 2. Evaluation metrics for general-purpose texture descriptors and radiomic features classified by SVM.

	Descriptors	Evaluation metrics				
		ACC (%)	PREC (%)	SEN (%)	F1 (%)	SPEC (%)
Texture descriptors	GLCM	76.40 ± 1.20	63.40 ± 1.80	64.10 ± 3.00	**63.70 ± 2.30**	82.30 ± 0.80
	LBP	73.00 ± 0.80	71.20 ± 4.20	28.20 ± 0.90	40.30 ± 0.10	94.50 ± 1.20
	LPQ	75.80 ± 1.70	72.70 ± 3.70	40.50 ± 4.90	51.90 ± 4.70	92.70 ± 1.10
Radiomic descriptors	FO	65.80 ± 0.80	47.60 ± 1.10	59.80 ± 4.50	53.00 ± 2.50	68.60 ± 1.10
	GLDM	70.30 ± 1.30	53.70 ± 1.80	60.80 ± 2.70	57.00 ± 2.10	74.90 ± 1.20
	GLRLM	72.90 ± 0.90	57.40 ± 1.30	63.30 ± 2.70	60.20 ± 1.50	77.50 ± 1.40
	GLSZM	64.40 ± 2.00	45.40 ± 3.00	53.50 ± 8.50	49.00 ± 5.30	69.90 ± 1.40
	NGTDM	66.90 ± 1.80	49.20 ± 1.90	64.90 ± 2.20	55.90 ± 1.10	67.80 ± 3.40
	SHAPE2D	87.00 ± 0.80	78.40 ± 1.50	82.90 ± 1.80	**80.50 ± 1.20**	89.00 ± 1.00

Table 3. Evaluation metrics for top late-fusion ensemble models and cross-dataset analysis. The combination of transfer learning and handcrafted features was developed in the BUS-BRA dataset. The late-fusion ensemble was evaluated through cross-dataset analysis on the BUSI and BUID datasets. Models with * indicate comparison with state-of-the-art systems.

Features	Dataset	Evaluation metrics				
		ACC (%)	PREC (%)	SEN (%)	F1 (%)	SPEC (%)
GLCM, GLDM, GLRLM, NGTDM, SHAPE2D, ResNet50, InceptionV3, EfficientNetB4, VGG16 (Ours)	BUS-BRA	89.90 ± 0.50	85.70 ± 0.30	82.20 ± 1.50	83.90 ± 0.90	93.50 ± 0.10
	BUSI	93.40	99.40	80.00	88.70	99.80
	BUID	77.60	80.90	75.60	78.20	79.80
HOG, LBP, shape, DenseNet201 [9]*	BUSI	96.00	96.00	96.00	96.00	96.00
HOG, ResNet50 [14]*	BUSI	88.87	87.50	86.90	87.18	83.80
GLCM, HOG, LBP, ResNet50 [15]*	BUSI	85.76	85.16	83.95	84.28	-
HoVer-Trans [17]*	BUSI	85.50	87.60	86.70	87.20	83.80

5. Conclusion

The aim of this research was to develop a binary classification system for breast lesions in ultrasound images by combining models utilizing transfer learning and handcrafted features. Among the transfer learning scenarios evaluated, ResNet50 stood out, with an F1-score of 81.97% for Input C images preprocessing. Among the handcrafted features, SHAPE2D stood out with an F1-score of 80.50%. The most effective combination of models using the late-fusion ensemble included all four transfer learning features and the GLCM, GLDM, GLRLM, NGTDM, and SHAPE2D descriptors, with an F1-score of 83.90% in the BUS-BRA dataset. In the cross-dataset evaluation, the top late-fusion ensemble model in the development dataset scored an F1-score of 88.70% and 78.20% in the test BUSI and BUID datasets, respectively. Therefore, this study has demonstrated that combining transfer learning and handcrafted features is complementary and enhances performance compared to using only a single feature. Moreover, the performance of the optimal combinations of the late-fusion ensemble was robust in a cross-dataset analysis. These results underscore the potential robustness of a late-fusion ensemble utilizing CNN transfer learning and handcrafted features to classify breast lesions in ultrasound images. For future research, we plan on improving classification performance by incorporating other medical images, such as ultrasound elastography. Another research direction deals with exploring explainable models that are able to highlight relevant portions in the image to aid the radiologist in the diagnostic.

Acknowledgements

This project was supported by the Ministry of Science, Technology and Innovation, with resources from Law No. 8,248, of October 23, 1991, within the scope of PPI-SOFTEX, coordinated by Softex and published PDI 03, DOU 01245.023862/2022-14.

Author contributions

Vanessa K. Foleis, Bianca A. Andrade, and Henrique B. Shigihara contributed to the conceptualization, model implementation, result analysis, and drafting of the manuscript. Dimas A. M. Lemes and José G. Picolo contributed to result analysis and manuscript review. Bernardo F. Junqueira, Guilherme R. Sales, Valentino Corso, and Cides S. Bezerra contributed to the conceptualization, supervision, result analysis, and manuscript review.

References

[1] ŁUKASIEWICZ, S. et al. Breast cancer—epidemiology, risk factors, classification, prognostic markers, and current treatment strategies—an updated review. *Cancers*, v. 13, n. 17, p. 4287, 2021.

[2] WANG, Y.; YAO, Y. Breast lesion detection using an anchor-free network from ultrasound images with segmentation-based enhancement. *Scientific Reports*, v. 12, n. 1, p. 14720, 2022.

[3] HOOLEY, R. J.; SCOUTT, L. M.; PHILPOTTS, L. E. Breast ultrasonography: State of the art. *Radiology*, v. 268, n. 3, p. 642–659, 2013.

[4] ISLAM, M. T.; TASCIOTTI, E.; RIGHETTI, R. Estimation of vascular permeability in irregularly shaped cancers using ultrasound poroelastography. *IEEE Transactions on Biomedical Engineering*, v. 67, n. 4, p. 1083–1096, 2020.

[5] GREENSPAN, H.; GINNEKEN, B. van; SUMMERS, R. M. Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Transactions on Medical Imaging*, v. 35, n. 5, p. 1153–1159, 2016.

[6] AKKUS, Z. et al. A survey of deep-learning applications in ultrasound: Artificial intelligence-powered ultrasound for improving clinical workflow. *Journal of the American College of Radiology*, v. 16, n. 19, p. 1546–1440, 2019.

[7] ZOURHRI, M. et al. Deep learning technique for classification of breast cancer using ultrasound images. In: *2023 3rd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*. Mohammedia, Marocco: IEEE, 2023. p. 1–8.

[8] ARDABILI, S.; MOSAVI, A.; FELDE, I. Advances of deep learning in breast cancer modeling. In: *2023 IEEE 21st Jubilee International Symposium on Intelligent Systems and Informatics (SISY)*. Pula, Croatia: IEEE, 2023. p. 501–508.

[9] CRUZ-RAMOS, C. et al. Benign and malignant breast tumor classification in ultrasound and mammography images via fusion of deep learning and handcraft features. *Entropy*, v. 25, n. 7, p. 991, 2023.

[10] AL-DHABYANI, W. et al. Dataset of breast ultrasound images. *Data in Brief*, v. 28, p. 104863, 2020.

[11] ARDAKANI, A. A. et al. An open-access breast lesion ultrasound image database: Applicable in artificial intelligence studies. *Computers in Biology and Medicine*, v. 152, p. 106438, 2023.

[12] ALI, M. D. et al. Breast cancer classification through meta-learning ensemble technique using convolution neural networks. *Diagnostics*, v. 13, n. 13, p. 2242, 2023.

[13] MOON, W. K. et al. Computer-aided diagnosis of breast ultrasound images using ensemble learning from convolutional neural networks. *Computer Methods and Programs in Biomedicine*, v. 190, p. 105361, 2020.

[14] ABHISHEKA, B. et al. Combining handcrafted and cnn features for robust breast cancer detection using ultrasound images. In: *2023 3rd International Conference on Innovative Sustainable Computational Technologies (CISCT)*. New Delhi, India: IEEE, 2023. p. 1–6.

[15] ABHISHEKA, B. et al. The efficacy of combining deep and handcrafted features for breast cancer classification using ultrasound images. In: *2023 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*. India: IEEE, 2023. v. 10, p. 735–740.

[16] NANNI, L. et al. Digital recognition of breast cancer using takhisisnet. In: ____. [S.l.]: IGI Global, 2022. p. 1286–1304.

[17] MO, Y. et al. Hover-trans: Anatomy-aware hover-transformer for roi-free breast cancer diagnosis in ultrasound images. *IEEE Transactions on Medical Imaging*, v. 42, n. 6, p. 1696–1706, 2023.

[18] GóMEZ-FLORES, W.; GREGORIO-CALAS, M. J.; PEREIRA, W. C. D. A. BUS-BRA: A breast ultrasound dataset for assessing computer-aided diagnosis systems. *Medical Physics*, p. 1–14, 2023.

[19] MAYERHOEFER, M. E. et al. Introduction to radiomics. *Journal of Nuclear Medicine*, v. 61, n. 4, p. 488–495, 2020.

[20] DHEEPAK, G.; J., A. C.; VAISHALI, D. Brain tumor classification: a novel approach integrating GLCM, LBP and composite features. *Frontiers in Oncology*, v. 13, p. 1248452, 2024.

[21] HALL-BEYER, M. Glcm texture: a tutorial. *National Council on Geographic Information and Analysis Remote Sensing Core Curriculum*, University of Calgary Calgary, AB, Canada, v. 3, n. 1, p. 75, 2000.

[22] OJALA, T.; PIETIKAINEN, M.; MAENPAA, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 24, n. 7, p. 971–987, 2002.

[23] OJANSIVU, V.; HEIKKILÄ, J. Blur insensitive texture classification using local phase quantization. *Image and Signal Processing*, Springer Berlin Heidelberg, Berlin, Heidelberg, p. 236–243, 2008.

[24] RAO, K. S. et al. Intelligent ultrasound imaging for enhanced breast cancer diagnosis: Ensemble transfer learning strategies. *IEEE Access*, v. 12, p. 22243–22263, 2024.

[25] KITTLER, J. et al. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 20, n. 3, p. 226–239, 1998.

[26] GESNOUIN, J. et al. Assessing cross-dataset generalization of pedestrian crossing predictors. In: *2022 IEEE Intelligent Vehicles Symposium (IV)*. [S.l.: s.n.], 2022. p. 419–426.