Ensemble Deep Learning Techniques for Advancing Breast Cancer Detection and Diagnosis

Adam M.Ibrahim[1], Jianqiang Li[1], and Yan Pei[2]

[1] Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China.
adam2222mohmd@gmail.com, lijianqiang@bjut.edu.cn

[2] Computer Science Division, University of Aizu, Aizuwakamatsu, Fukushima, 965-8580, Japan.
Email: peiyan@u-aizu.ac.jp

Abstract. The integration of deep learning (DL) and digital breast tomosynthesis (DBT) presents a unique opportunity to improve the reliability of breast cancer (BC) detection and diagnosis while accommodating novel imaging techniques. This study utilizes the publicly available Mammographic Image Analysis Society (MIAS) database v1.21 to evaluate DL algorithms in identifying and categorizing cancerous tissue. The dataset has undergone preprocessing and has been confirmed to be of exceptional quality. Transfer learning techniques are employed with three pre-trained models - Mobilenet, Xception, Densnet, mobilenet lstm - to improve performance on the target task. Stacking Ensemble learning techniques will be utilized to combine the predictions of the best-performing models to make the final prediction for the presence of BC. The evaluation will measure the performance of each model using standard evaluation metrics, including accuracy (ACC), precision (PREC), recall (REC), and F1-score (F1-S). This study highlights the potential of DL in enhancing diagnostic imaging and advancing healthcare.

Keywords: Breast Cancer, Deep Learning, Ensemble Learning, Detection

1 Introduction

BC is a frequent and lethal illness, making risk prediction difficult. Mammography is the most expensive early detection technology, and a standardized and community-based screening approach has been proposed to address this [1, 2]. Cancer risk prediction methods use multiple risk factors, such as molecular genetics, imaging, and public health data, to accurately predict the likelihood of BC based on individual diagnostic imaging screenings [3]. Breast density is not a reliable predictor of BC risk, as it is used to determine the frequency of screening [4]. Mammography screening is essential to reduce death rates from breast cancer, but age is the main factor used to select people for screening. Interest is growing in customized screening methods [5]. Risk stratification using disease prediction models can identify women at risk of developing BC, allowing tailored surveillance to maximize benefit [6]. A technique used in histopathology photos to find cancer is the BC detection factor [7]. Cancer risk models are used to assess cancer risk and project outcomes, based on the elevated risk of BC linked to various characteristics, without any connection to the type of mammography used [8].

Cancer is a major global public health issue, increasing prevalence in industrialized and developing countries [9]. Breast disease is the unchecked, potentially cancerous proliferation of breast cells, and microscopic histopathological examinations depend on expert visual interpretation, which is subjective and dependent on the observer's knowledge [10]. The process of multi-classification cancer diagnosis utilizing histology images is difficult and time-consuming due to the lack of qualified pathologists in many low-income nations. This can lead to incorrect findings due to the intricacy of the pictures and the pathologist's limited ability to comprehend a large quantity of data [11, 12]. Misinterpretation of screening mammography can lead to overdiagnosis, costing people money. To increase the efficacy of screening mammography, a new methodology was developed, combining picture characteristics and a forecasting technique. Bidirectional screening mammography density imbalance was used as a signal to assess the likelihood of BC in computed tomography images [13]. The experiment tested whether a DL-based method could outperform established frameworks for identifying cancer risk, as patients often have repeated mammography examinations during BC monitoring [14]. Predicting the results of a single abnormal mammogram is the screening task, but we did not use many priors as inputs to the models. [15]. Research opportunities for biological-subtype intelligent forecasting have become available due to the rapid development of BC detection technologies, but forecasting for biological subgroups is still a difficult problem.

1.1 Related Work

According to the ratings of the remaining data, writers in [16] The authors used a deep feed-forward network to train a RankDeepSurv model to predict relapse in patients with nasopharyngeal cancer. The C-index of the RankDeepSurv model was higher than more conventional survival analysis techniques, with a C-index of 0.681. Using images of tumor cell extracts, the authors of [17] The scientists present a DL network that blends convolutional neural networks (CNNs) with recurrent models to predict the prognosis of colorectal cancer. They examined 420 colorectal cancer patient tumor samples and found that DL algorithms can derive more predictive information from tissue shape than conventional human observation techniques. In [18], DeepSurv is a deep neural network (DNN) for survival analysis based on Cox regression hazard models. It predicts the correlation between a person's variables and clinical result, using link weights to determine how a patient's variables impact their level of risk. It outperforms existing advanced survival models and predicts more intricate relationships between a participant's characteristics and failure risk.

Without the help of a toxicologist to pinpoint specific areas, MesoNet is a deep convolutional neural network method that estimates the likelihood that mesothelioma patients will survive, according to research published in [19]. MesoNet can identify regions linked to patient outcomes, and researchers found that these regions are usually found in the stroma and histologically. The study's findings suggest that DL algorithms may be able to identify previously unknown features that are predictive of clinical results, resulting in the discovery of novel biomarkers. Researchers outline three techniques in [20]. A solitary training batch is used to assess the effectiveness of training CNNs. Deep layer activations are subjected to a dispersed stochastic neighbor modeling approach to show how the classes are separated. Finally, DeepDream with special settings such as pyramid level 12, 75 iterations, a scale of 1.1, and histogram stretching is used to show the activation of deep neurons in the 46 layers of the VGG19 DL model. Without requiring a separate tumor segmentation step, scientists in [21] have presented three residual DNN models as options to estimate methylation conditions. According to the study, ResNet50 outperformed ResNet18 and ResNet34 with a statistically significant ACC of 94.90%.

2 Methodology

This section will outline our approach, broken down into steps illustrated in Figure. 1.

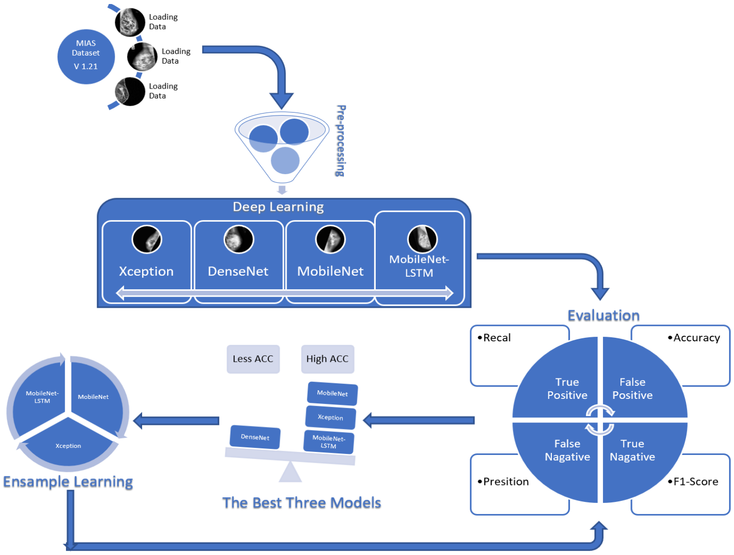

Fig. 1. The proposed framework for breast cancer classification involves collecting mammography images from the MIAS dataset, comparing the performance of each model, identifying the best-performing models for feature extraction, and training multiple models using ensemble learning. The predictions are then pooled to create a final forecast.

2.1 Dataset

The Mammographic Image Analysis Society (MIAS) database v1.21 has been used to identify and diagnose breast cancer. It contains 322 digitized mammograms, of which

208 have benign and 114 have malignant labels. The ground truth data includes the size and shape of any masses or microcalcifications, the level of speculation, and the existence of architectural distortions. These data are important for understanding the normal structure of the breast tissue.

2.2 Preprocessing

Preprocessing steps are essential for medical image analysis tasks, as they significantly impact the model's performance. Inaccurate preprocessing can introduce image artifacts or anomalies, potentially compromising the model's ACC. It is important to ensure that the chosen preprocessing methods are appropriate for the specific task and that the data is of exceptional quality. The dataset has already undergone preprocessing and is available online. However, it is still important to confirm that the applied preprocessing techniques are suitable for the task and that the preprocessing process has not introduced any artifacts or errors.

2.3 Deep Learning

This section will elaborate on the DL techniques utilized in our approach.

MobileNet MobileNet architecture was implemented, consisting of convolutional and pooling layers, followed by multiple fully connected layers. ReLU activation function was used, running 100 epochs and a batch size of 16, Adam optimizer, categorical cross-entropy loss function and ACC metric for evaluation. To fine-tune the pretrained model, the weights of the convolutional layers were frozen and only trained the fully connected layers.

Xception The pre-trained Xception model was fed a predetermined input shape and size, with the include top parameter set to False. Fresh layers were added, including dropout layers to avoid overfitting and fully linked layers with ReLU activation functions. Two nodes in the output layer have a softmax activation function. The model was trained for 100 iterations, with the training data being randomly mixed before each iteration.

MobileNet-LSTM This approach utilized the MobileNet architecture as a feature extractor and added an LSTM layer for sequence processing. The model was built using the Adam optimizer with a batch size of 16, and trained for 100 iterations, with the training data being randomly mixed before each iteration. To incorporate sequence processing, the output of the fully connected layers was reshaped into a 3D tensor with a shape of (batch size, time steps, input dim). The LSTM layer had 256 units and dropout of 0.3, and a softmax output layer with two nodes was added. The model was trained for 100 iterations, with the training data being randomly mixed before each iteration.

DenseNet The DenseNet201 model is pre-trained on the ImageNet dataset, and the last few layers are replaced with new layers for our specific classification task. The first few lines of code load the pre-trained DenseNet201 model with the input shape of the images, exclude the top layers and set the pooling method to average. Then, we define the new layers on top of the pre-trained model, taking inputs from the pre-trained model and outputs from the new layers. We compile the model using the Adam optimizer, categorical cross-entropy loss, and ACC metric and fit the model on the training data for 100 epochs with a batch size of 16 and shuffle the data after each epoch.

2.4 Ensemble Learning

After comparing the performance of the DL, we will select the best three models based on their evaluation metrics. We will then use an ensemble learning tech-nique, such as voting or stacking, to combine the predictions of these three models to make a final prediction for the presence of BC in mammography images. The ensemble learning technique can help improve the model's overall performance by combining the strengths of each individual model and reducing their weaknesses.

2.5 Evaluation

During the assessment phase, common evaluation metrics such as ACC, PREC, REC, and F1-S will be used to assess each model's performance. These metrics are calculated using the confusion matrix, which lists a classification model's performance in four categories: true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). ACC measures the fraction of properly classified samples, PREC measures the proportion of correctly classified positive samples over the total number of positive predictions, REC measures the proportion of correctly classified positive samples over the total number of positive samples in the dataset, and F1-S is a harmonic mean of PREC and REC. These metrics are useful for datasets with imbalanced classes and can be calculated using specific formulas, as depicted in Eqs. 1-4.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$PREC = \frac{TP}{TP + FP} \tag{2}$$

$$REC = \frac{TP}{TP + FN} \tag{3}$$

$$F1 - S = 2 \cdot \frac{PREC \cdot REC}{PREC + REC} \tag{4}$$

3 Results and Comparison

3.1 Results

The results show that MobileNet achieved the highest ACC of 93.34%, followed by MobileNet-LSTM with an ACC of 92.64%. Xception achieved an ACC of 88.62%, and Densnet achieved an ACC of 87.23%. Looking at the confusion matrix Figure 2 of each model, we can see that MobileNet had the fewest misclassifications, with only 18 false negatives and 30 false positives. Xception had more false positives, with 52 misclassifications, while Densnet had the same number of false positives but more false negatives. It's important to note that these results were obtained using the same dataset and training procedure, so the differences in performance can be attributed to the architecture of the models. MobileNet and MobileNet-LSTM both use a lightweight architecture optimized for mobile devices, which may have contributed to their superior performance. Xception, on the other hand, is a deeper and more complex model, which may have made it more difficult to train effectively with the limited dataset. Densnet has similar layers to Xception but a different architecture, which may have contributed to its lower performance. Additionally, although we used transfer learning to initialize the models' weights with pre-trained weights from ImageNet, the particular pre-trained model used may have affected the model's performance. The results suggest that the MobileNet architecture is well-suited for this classification task and outperforms other architectures, such as Xception and Densnet. However, further experimentation with different architectures and datasets may yield different results. Multiple models are trained using the ensemble learning approach, and then their predictions are pooled to get a final forecast. We have chosen the three best-performing models, MobileNet, MobileNet-LSTM, and Xception, to create an ensemble model. The ensemble model combines the predictions of three models using a simple voting scheme. The class with the most votes is considered the final prediction. The ensemble model has achieved an ACC of 94.45%, which is a significant improvement over the individual models' performances. Comparing the confusion matrices of the individual models and the ensemble model, it can be seen that the ensemble model has fewer misclassifications due to the fact that it has taken into account the strengths and weaknesses of each model and made a final prediction based on their combined expertise. In conclusion, the ensemble model has achieved the highest ACC, and it can be seen that combining multiple models' predictions has led to a more accurate and robust model. The classification report shows that all the models have achieved high ACC, PREC, REC, and F1-S in detecting malignant and benign tumors. The MobileNet and MobileNet-LSTM models achieved the highest PREC scores of 0.93 and 0.91 for benign tumors, respectively, while the ensemble model achieved the highest PREC score of 0.95 for malignant tumors. The MobileNet-LSTM model achieved the highest REC score of 0.96 for benign tumors, while the Xception model achieved the highest REC score of 0.91 for malignant tumors. The Densenet model achieved the lowest ACC and F1-S among the individual models. From Figure 3, The MobileNet-LSTM model was trained for 100 epochs, with a batch size of unspecified size. The training ACC started at 60.12% in the first epoch and steadily increased to 98.92% by the end of the training. The loss

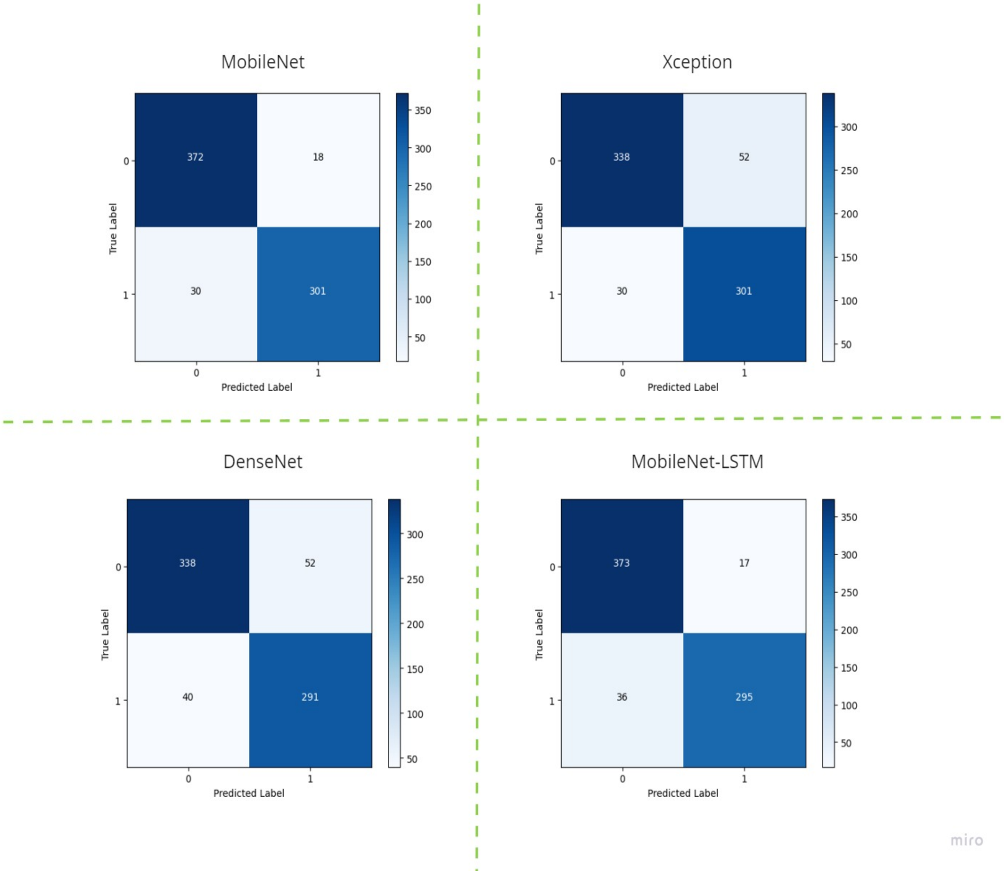

Fig. 2. MobileNet had the fewest misclassifications, with 18 false negatives and 30 false positives. Xception had more false positives, Densnet had the same number of false positives but more false negatives, and MobileNet-LSTM had 17 false negatives and 36 false positives.

function decreased from 0.7559 to 0.0270 over the course of the training. For Xception, the ACC increased steadily from 55.81% to 81.50%, and then experienced more fluctuations in ACC but continued to improve overall, reaching a maximum ACC of 94.59% by the 30th epoch. After that, the model's ACC seemed to plateau and remain relatively stable, hovering around 93-94%. The training process seems to have been successful in producing a well-performing model. The results represent the performance of four different models Figure 4, namely MobileNet, Xception, DenseNet, and MobileNet-LSTM, evaluated using the receiver operating characteristic (ROC) curve analysis. The ROC curve is a graphical representation of the trade-off between sensitivity (true positive rate) and specificity (true negative rate) of a binary classifier as the decision threshold is varied. The ROC curve for MobileNet has an AUC of 0.99, indicating excellent performance in distinguishing between positive and neg-

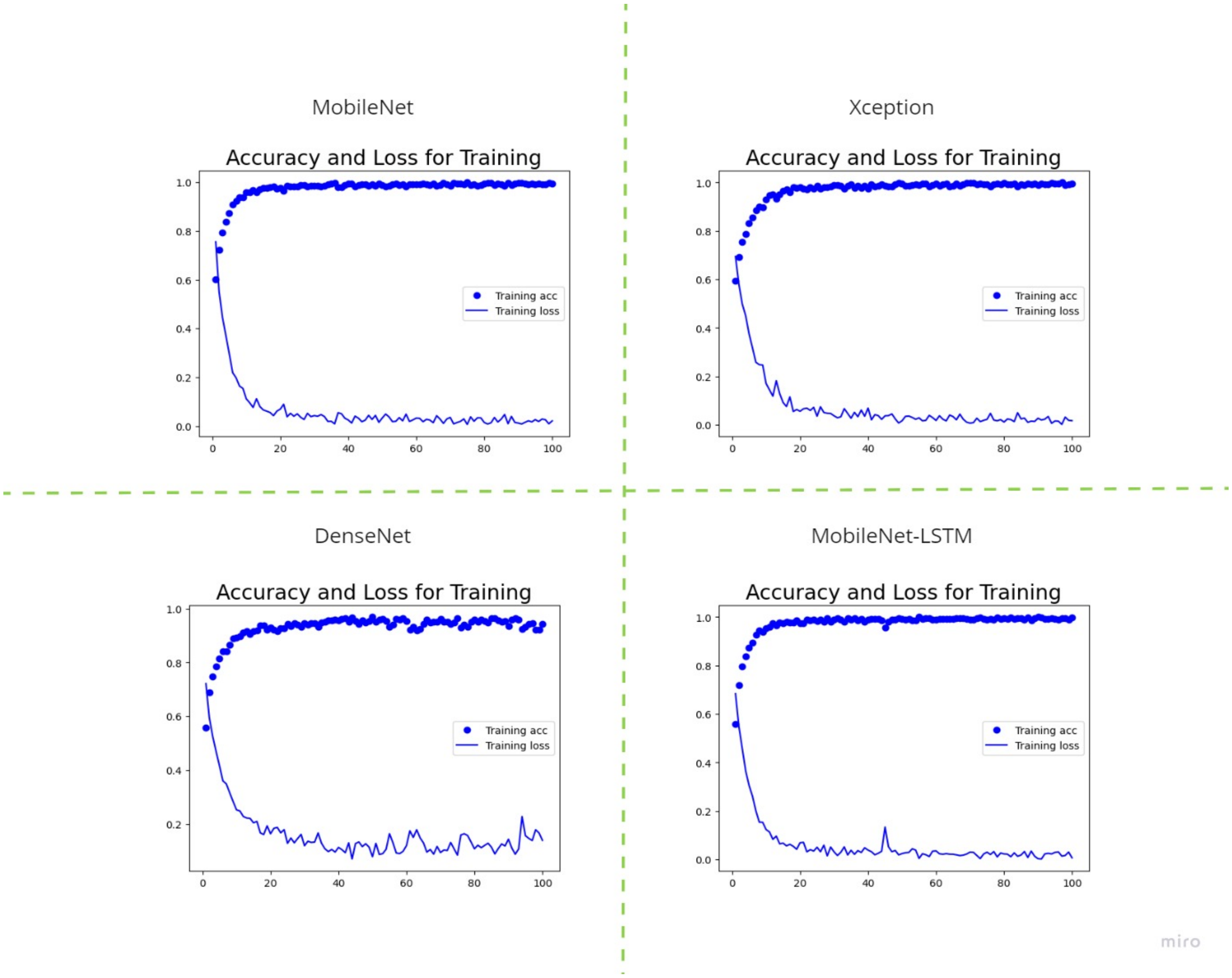

Fig. 3. The proposed algorithms improved their training and testing accuracy with each epoch, and after 100 epochs, the hybrid Model MobileNet-LStm achieved 0.99 accuracies. The training cross-entropy loss function decreased with each epoch, with a minimum decrease in the final epoch.

ative classes. Xception has an AUC of 0.96, DenseNet has an AUC of 0.93, and the MobileNet-LSTM model has an AUC of 0.98, similar to Xception. The results suggest that MobileNet and MobileNet-LSTM models outperform the other two models in terms of AUC, while all models perform well in distinguishing between two classes, with DenseNet having the worst performance with an AUC of 0.93. These results are important when selecting a suitable model for a given task, as the trade-off between performance and computational complexity needs to be carefully considered. In Table

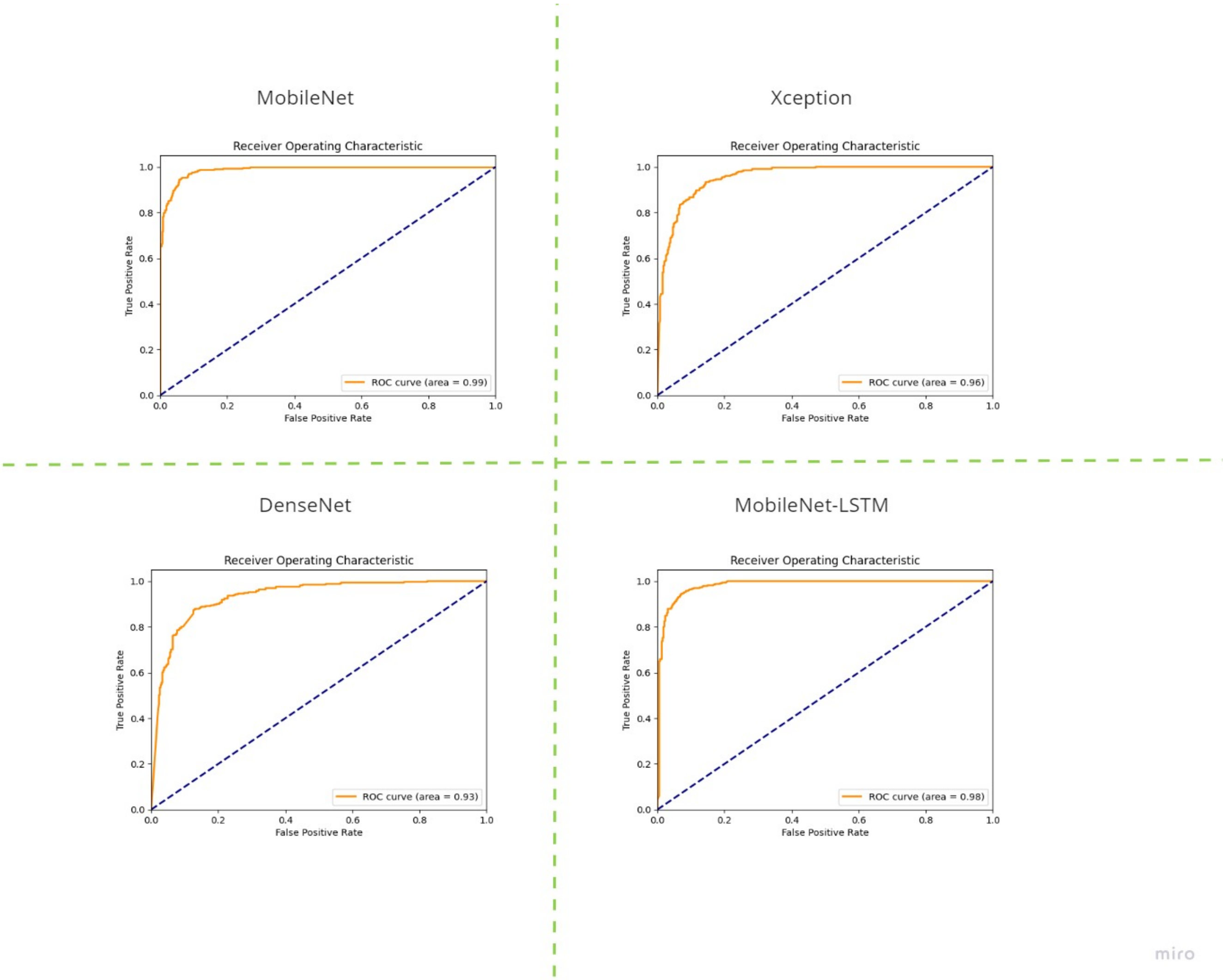

Fig. 4. The hybrid MobileNet-LSTM model achieves an AUC of 0.98, indicating excellent classification performance and fewer false-negative results than standard CNN-based models.

1, this study evaluated five models with performance metrics such as ACC, PREC, REC, F1-S, and ROC AUC. The ensemble model had the highest ACC score with a score of 94.45%, followed by MobileNet with a score of 93.34%, MobileNet-LSTM with a score of 92.64%, Xception with a score of 88.62%, and DenseNet with a score of 87.23%. The ensemble model also had the highest PREC, REC, and F1-S, as explained in Table 2. The ROC AUC was not available for the ensemble model as it is

Table 1. Summary of results for MobileNet, Xception, DenseNet, MobileNet-LSTM, and Ensemble models.

Method	ACC	PREC	REC	F1-S	Roc Auc
MobileNet	0.9334	0.9300	0.9500	0.9400	0.9900
Xception	0.8862	0.9200	0.8700	0.8900	0.9600
DenseNet	0.8723	0.8900	0.8700	0.8800	0.9300
MobileNet-LSTM	0.9264	0.9100	0.9600	0.9300	0.9800
Ensemble	0.9445	0.9400	0.9600	0.9500	-

Table 2. Summary of cancer prediction studies using DL models

Article	Cancer Type	Methodology	Result/Performance
[16]	BC	Deep neural network	Cancer prediction, C-index 0.704
[17]	Colorectal cancer	VGG16	1Cancer diagnosis, HR 2.3, CI 95 percent 1.79-3.03, AUC 0.69
[18]	BC	Neural Network	Cancer prediction, CI 0.67
[19]	BC	CNNs	Cancer prediction, ACC 87%
[20]	Colorectal cancer	VGG19, GoogLeNet, Resnet50, AlexNet, SqueezeNet	CI 95 classification of 9 tissues
[21]	Glioblastoma multiforme	Deep neural network	Cancer prediction, ResNet50 : 94.90% (+/-3.92%); ResNet34 (34 layers) : 80.72% (+/-13.61%)
Our Study	Breast cancer	MobileNet, Xception, DenseNet, MobileNet-LSTM and Ensemble Learning	Ensemble Learning :94.54%

not a binary classifier. Overall, the ensemble model outperformed the other models in terms of ACC and other performance metrics.

4 Conclusion

This study demonstrates the potential of digital breast tomosynthesis (DBT) to enhance BC detection and diagnosis. Four pre-trained models - Mobilenet, Xception, Densnet, and Mobilenet-lstm - have shown promising results in identifying and categorizing cancerous tissue. The ensemble model, which combines the predictions of the best-performing models, achieved the highest ACC and outperformed all individual models. These findings suggest that transfer learning and ensemble learning techniques can be used to improve the reliability of BC detection and diagnosis, while accommodating novel imaging techniques.

5 Future Work

Future research in medical imaging and AI for breast cancer detection and diagnosis could include exploring larger datasets, incorporating clinical and patient-specific data, and evaluating proposed models in clinical settings. Additionally, investigating the potential of DL models for predicting treatment response and recurrence risk could inform personalized treatment plans. Ethical and regulatory frameworks are needed to ensure responsible and safe integration of DL models into clinical practice.

Funding

This study is supported by the National Key R&D Program of China with the project no. 2020YFB2104402.

Conflict of interest

The authors declare that they have no Conflict of interest.

Acknowledgments

The authors thank China's National Key R&D Program for providing the experimental facilities to perform these experiments.

References

1. Syed Hamza Shah, Muhammad Javed Iqbal, Irfan Ahmad, Saif Khan, and Joel JPC Rodrigues. Optimized gene selection and classification of cancer from microarray gene expression data using deep learning. *Neural Computing and Applications*, 32(22):17457–17468, 2020.
2. Wael Gouda, Mohammad Almurafeh, Muhammad Humayun, and Nadeem Zia Jhanjhi. Detection of covid-19 based on chest x-rays using deep learning. *Healthcare*, 10(4):343, 2022.
3. Saad Abdulazeez Ismael, Ammar Mohammed, and Hossam Hefny. An enhanced deep learning approach for brain cancer mri images classification using residual networks. *Artificial Intelligence in Medicine*, 102:101779, 2020.
4. Noureddine Dif and Zakaria Elberrichi. A new deep learning model selection method for colorectal cancer classification. *International Journal of Swarm Intelligence Research*, 11(2):72–88, 2020.
5. Shah Nawaz Brohi, Tiju R Pillai, Nisar N Brohi, and Nadeem Zia Jhanjhi. A multilayer perceptron model for the classification of breast cancer cells. *International Journal of Computing and Digital Systems*, 10(2):104–115, 2021.
6. Aditi Khamparia, Prerna K Singh, Poonam Rani, Debasis Samanta, Aditya Khanna, and Bharat Bhushan. An internet of health things-driven deep learning framework for detection and classification of skin cancer using transfer learning. *Trans. Emerg. Telecommun. Technol.*, 32:e3963, 2021.
7. Roshan A Welikala, Paolo Remagnino, Jen Hong Lim, Chee Seng Chan, Srikumar Rajendran, Thomas George Kallarakkal, Rosnah Binti Zain, Ruwan D Jayasinghe, Jyoti Rimal, Alexander R Kerr, et al. Automated detection and classification of oral lesions using deep learning for early detection of oral cancer. *IEEE Access*, 8:132677–132693, 2020.

8. Muhammad Humayun and Ahmed Alsayat. Prediction model for coronavirus pandemic using deep learning. *Comput. Syst. Sci. Eng.*, 40:947–961, 2022.
9. I Pacal, D Karaboga, A Basturk, B Akay, and U Nalbantoglu. A comprehensive review of deep learning in colon cancer. *Comput. Biol. Med.*, 126:104003, 2020.
10. Ghulam Murtaza, Liyana Shuib, Ainuddin Wahid Abdul Wahab, Ghulam Mujtaba, Henry Friday Nweke, Mohammed Ali Al-Garadi, Faisal Zulfiqar, Gohar Raza, and Nurul Aini Azmi. Deep learning-based breast cancer classification through medical imaging modalities: State of the art and research challenges. *Artif. Intell. Rev.*, 53:1655–1720, 2020.
11. Wei Chi, Long Ma, Jinglei Wu, Ming Chen, Wenjing Lu, and Xianfeng Gu. Deep learning-based medical image segmentation with limited labels. *Phys. Med. Biol.*, 65:235001, 2020.
12. Rong Qin, Zhiqiang Wang, Liying Jiang, Kai Qiao, Jie Hai, Jiaojiao Chen, Jinyi Xu, Dongdong Shi, and Bin Yan. Fine-grained lung cancer classification from pet and ct images based on multidimensional attention mechanism. *Complexity*, 2020:6153657, 2020.
13. Rohit Manne, Siva Kantheti, and Srilekha Kantheti. Classification of skin cancer using deep learning, convolutional neural networks-opportunities and vulnerabilities-a systematic review. *International Journal of Modern Trends in Science and Technology*, 01(12):2455–3778, 2020.
14. Hyun-Seok Shon, Enkhchimeg Batbaatar, Kyung-Ok Kim, Eun-Jae Cha, and Kyung-A Kim. Classification of kidney cancer data using cost-sensitive hybrid deep learning approach. *Symmetry*, 12(1):154, 2020.
15. Hyun-Seok Shon, Enkhchimeg Batbaatar, Kyung-Ok Kim, Eun-Jae Cha, and Kyung-A Kim. Automated detection and classification of oral lesions using deep learning to detect oral potentially malignant disorders. *Cancers*, 13(11):2766, 2021.
16. Bing Jing, Ting Zhang, Zhe Wang, Yan Jin, Ke Liu, Wei Qiu, Liang Ke, Yuhao Sun, Chuan He, Dongdong Hou, et al. A deep survival analysis method based on ranking. *Artificial Intelligence in Medicine*, 98:1–9, 2019.
17. Dmitri Bychkov, Nina Linder, Riku Turkki, Sten Nordling, Panu E. Kovanen, Clare Verrill, Maria Walliander, Johan Lundin, Caj Haglund, and Mikael Lundin. Deep learning based tissue analysis predicts outcome in colorectal cancer. *Scientific Reports*, 8:3395, 2018.
18. Jared L. Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: Personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1):24, 2018.
19. Saud Al Alanazi, Joarder Mohammad Mamun Kamruzzaman, Nuzhat Islam Sarker, Moudhi Alruwaili, Yahya Alhwaiti, Noura Alshammari, and Mohammad Hassan Siddiqi. Boosting breast cancer detection using convolutional neural network. *Journal of Healthcare Engineering*, 2021:5528622, 2021.
20. Jakob Nikolas Kather, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Esther Herpel, Cleo-Ann Weis, Timo Gaiser, Alexander Marx, Nektarios Athanasios Valous, Dieter Ferber, et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS Medicine*, 16(1):e1002730, 2019.
21. Panagiotis Korfiatis, Timothy L. Kline, Daniel H. Lachance, Ian F. Parney, Jan C. Buckner, and Bradley J. Erickson. Residual deep convolutional neural network predicts MGMT methylation status. *Journal of Digital Imaging*, 30(5):622–628, 2017.