Article

Developing a supplementary diagnostic tool for breast cancer risk estimation using ensemble transfer learning

Tengku Muhammad Hanis[1], Nur Intan Raihana Ruhaiyem[2], Wan Nor Arifin[3], Juhara Haron[4,5], Wan Faiziah Wan Abdul Rahman[5,6], Rosni Abdullah[2], Kamarul Imran Musa[1,*]

[1] Department of Community Medicine, School of Medical Sciences, Universiti Sains Malaysia, Kubang Kerian 16150, Kelantan, Malaysia; tengkuhanismokhtar@gmail.com (T.M.H.); drkamarul@usm.my (K.I.M.)
[2] School of Computer Sciences, Universiti Sains Malaysia, USM 11800, Penang, Malaysia; intanraihana@usm.my (N.I.R.R.); rosni@usm.my (R.A.)
[3] Biostatistics and Research Methodology Unit, School of Medical Sciences, Universiti Sains Malaysia, Kubang Kerian 16150, Kelantan, Malaysia; wnarifin@usm.my
[4] Department of Radiology, School of Medical Sciences, Universiti Sains Malaysia, Kubang Kerian 16150, Kelantan, Malaysia; drjuhara@usm.my
[5] Breast Cancer Awareness and Research Unit, Hospital Universiti Sains Malaysia, Kubang Kerian 16150, Kelantan, Malaysia; drjuhara@usm.my (J.H.); wfaiziah@usm.my (W.F.W.A.R.)
[6] Department of Pathology, School of Medical Sciences, Universiti Sains Malaysia, Kubang Kerian 16150, Kelantan, Malaysia; wfaiziah@usm.my
* Correspondence: drkamarul@usm.my (K.I.M); tengkuhanismokhtar@gmail.com (T.M.H.)

Abstract: This study utilised an ensemble of pre-trained networks and digital mammograms to develop a supplementary diagnostic tool for radiologists. Digital mammograms and their associated information were collected from the department of radiology and pathology, Hospital Universiti Sains Malaysia. Thirteen pre-trained networks were selected and explored in this study. ResNet101V2 and ResNet152 had the highest mean PR-AUC, MobileNetV3Small and ResNet152 had the highest mean precision, ResNet101 had the highest mean F1 score, and ResNet152 and ResNet152V2 had the highest mean Youden J index. Subsequently, three ensemble models were developed using the top three pre-trained networks based on PR-AUC, precision, and F1 score. The final ensemble model had a mean precision, F1 score, and Youden J index of 0.82, 0.68, and 0.12, respectively. Additionally, the final model demonstrated a balanced performance across mammographic density. In conclusion, this study exhibited the good performance of the ensemble transfer learning on digital mammograms for the purpose of breast cancer risk estimation. This model can be utilised as a supplementary diagnostic tool for radiologists, thus, reducing their workloads.

Keywords: Asian women; breast cancer; transfer learning; artificial intelligence

1. Introduction

Artificial intelligence (AI) is expected to improve the efficiency of the healthcare system, including the area of oncology and radiology. AI had been studied to be used with thoracic imaging, abdominal and pelvic imaging, colonoscopy, mammography, brain imaging, and radiation oncology [1]. Digital mammograms have been widely used as part of breast cancer assessment. The use of screening mammogram has been shown to improve the early detection of breast cancer, which in turn reduce breast cancer mortality [2]. The introduction of mammogram-related AI to assist radiologists in breast cancer assessment may reduce their workloads and further improve the diagnostic accuracy of the mammogram reading. Additionally, radiologists will be more available to engage and focus on more complex medical cases or higher-level tasks. In fact, AI had been shown to reduce the time required by radiologists to interpret the mammogram, thereby improving overall cancer detection [3].

Transfer learning or pre-trained network is a network previously trained on a large dataset [4]. The use of pre-trained networks is expected to reduce the training time and

 doi:10.20944/preprints202303.0015.v1

improve the overall performance of the deep learning task [5]. The early layer of the convolutional neural network (CNN) learns the general and broader aspects of the image such as edges, textures, and patterns, while the last few layers learn more specific features of the image related to the task [6]. Hence, the main idea of transfer learning is to transfer the early learned layers trained on one task to another. There were two approaches to implementing transfer learning: (1) feature extraction and (2) fine-tuning. The former allows the previously trained network to be used on a different task without the need to train from a scratch, while the latter allows some adjustments to the pre-trained network by unfreezing a few last layers. The fine-tuned approach allows the pre-trained network to adapt to the new task and may further improve its performance on the task.

Mammographic density or breast density is one of the important risk factors of breast cancer. Mammographic density reflects the amount of dense tissue in the breast. Several factors that had been shown to affect breast density were age, number of children, body mass index, and menopause status [7]. Moreover, women with Asian ancestry had been observed to have denser breasts compared to other ethnicities [8]. Additionally, studies had shown that breast density reduces the sensitivity of mammograms [9,10]. Thus, any mammogram-based diagnostic tool should consider this risk factor in developing the tool. This study aims to develop a supplementary diagnostic tool for radiologists. Thus, this study explores the use of ensemble transfer learning and digital mammograms in breast cancer risk estimation. Furthermore, the performance of the model will be evaluated across dense and non-dense breast cases.

2. Related works

Several studies had been conducted related to the use of transfer learning on digital mammograms for breast cancer classification. Saber *et al*. [11] explored the use of six pre-trained networks for breast cancer classification. The study managed to achieve an accuracy of 0.99 with VGG16 as the best-performed model. Another study published in the same year explored the use hybrid model by combining a modified VGG16 and ImageNet and managed to achieve an accuracy of 0.94 [12]. Several other studies had managed to achieve good performance with VGG16 and VGG19 as well [13–15]. In addition, a study by Guan & Loew [16] comparing the feature extraction and fine-tuning approach using VGG16 showed that the latter approach performed better compared to the former though the difference in performance was very minimal in their study.

Several studies had explored the use of ResNet for breast cancer detection. Yu & Wang [17] compared several ResNet models including ResNet18, ResNet50, and ResNet101 in their study. ResNet18 had the highest accuracy at 0.96 outperforming all other ResNet variants. Another study compared several pre-trained networks including ResNet50, NasNet, InceptionV3, and MobileNet [18]. Essentially, this study applied two different pre-processing approaches to the mammogram images. Otsu thresholding was not applied in the first approach but was applied in the second approach. ResNet50 was the best model in the first approach with an accuracy of 0.78, while NasNet was the best model in the second approach with an accuracy of 0.68.

Additionally, a study by Ansar [19] proposed a transfer learning network using a MobileNet architecture for breast cancer classification. This study utilised two datasets separately, namely the Digital Database for Screening Mammography (DDSM) and curated breast imaging subset of DDSM (CBIS-DDSM), and achieved an accuracy of 0.87 and 0.75, respectively. Therefore, the result of this study suggests the use of different datasets may influence the performance of the transfer learning model.

Furthermore, other pre-trained network architectures had been researched for the purpose of breast cancer classification using digital mammograms. Jiang *et al*. [20] compared transfer learning models and deep learning models training from a scratch and they also compared GoogleNet and AlexNet for breast cancer classification. The study reported that transfer learning and GoogleNet outperformed the other. Another study explored the utilisation of InceptionV3 architecture on the INBreast dataset and achieved the highest AUC at 0.91 [21]. Recently, a study by Pattanaik *et al*. [22] proposed a hybrid transfer

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 March 2023 doi:10.20944/preprints202303.0015.v1

learning model consisting of DenseNet121 and extreme learning machine (ELM). The model achieved an accuracy of 0.97 and outperformed other models in the study. Table 1 presents the summary of previous works related to pre-trained networks and breast cancer classification that utilised digital mammograms. Notably, all the aforementioned studies utilised publicly available datasets in their studies except for Mendel *et al*. [14].

Table 1. Summary of the previous works related to pre-trained networks and breast cancer classification that utilised digital mammograms.

Study	Database	Pre-trained network	Performance metrics[1]
Pattanaik2022 [22]	DDSM	VGG19, MobileNet, Xception, ResNet50V2, InceptionV3, InceptionResNetV2, DenseNet201, DenseNet121, DenseNet121 + ELM[2]	Accuracy = 0.97 Sensitivity = 0.99 Specificity = 0.99
Khamparia2021 [12]	DDSM	AlexNet, ResNet50, MobileNet, VGG16, VGG19, MVGG16, MVGG16, ImageNet[2]	Accuracy = 0.94 AUC = 0.93 Sensitivity = 0.94 Precision = 0.94 F1 score = 0.94
Sabeer2021 [11]	MIAS	Inception V3, InceptionV2, ResNet, VGG16[2], VGG19, ResNet50	Accuracy = 0.99 AUC = 1.00 Sensitivity = 0.98 Specificity = 0.99 Precision = 0.97 F1 score = 0.98
Ansar2020 [19]	DDSM CBIS-DDSM	AlexNet, VGG16, VGG19, ResNet50, GoogLeNet, MobileNetV1[2], MobileNetV2	Accuracy = 0.87 Sensitivity = 0.95 Precision = 0.84
Falconi2020 [15]	CBIS-DDSM	VGG16[2], VGG19, Xception, Resnet101, Resnet152, Resnet50	Accuracy = 0.84 AUC = 0.84 F1 score = 0.85
Falconi2019 [18]	CBIS-DDSM	MobileNet, ResNet50[2], InceptionV3, NasNet	Accuracy = 0.78
Guan2019 [13]	DDSM	VGG16[2]	Accuracy = 0.92
Mendel2019 [14]	Primary data	VGG19[2]	AUC = 0.81
Yu2019 [17]	Mini-MIAS	ResNet18[2], ResNet50, ResNet101	Accuracy = 0.96
Mednikov2018 [21]	INbreast	InceptionV3[2]	AUC = 0.91
Jiang2017 [20]	BCDR-F03	GoogLeNet[2], AlexNet	AUC = 0.88
Guan2017 [16]	MIAS DDSM	VGG16[2]	Accuracy = 0.91 AUC = 0.96

[1] Performance metrics of the best or final model in the study
[2] Model with best performance metrics/selected as the final model in the study
DDSM=digital database for screening mammography, MIAS=mammographic image analysis society, CBIS-DDSM=curated breast imaging subset of database for screening mammography, BCDR-F03=breast cancer digital repository-film mammography dataset number 3, ELM=extreme learning machine, MVGG16=modified VGG16

3. Materials and Methods

3.1. Data

Two data were utilised in this study. Digital mammograms and their reports were retrieved from the department of radiology, Hospital Universiti Sains Malaysia (HUSM), and histopathological examination (HPE) results were retrieved from the department of

 doi:10.20944/preprints202303.0015.v1

pathology, HUSM. Generally, each set of the mammogram may consist of the right and left sides. Each side may consist of mediolateral oblique and craniocaudal views. Additionally, the mammogram reports contained information on the Breast Imaging-Reporting and Data System (BI-RADS) breast density and classification, while the HPE results contained information on the classification of the breast lesion. The data was collected from 1st January 2014 until 30th June 2020 from each respective department. Next, the two data were combined if the HPE data was dated within a year after the date of the mammogram was taken.

BI-RADS breast density was used to classify the mammograms into non-dense and dense breasts. The non-dense breast cases consisted of BI-RADS density of A and B, while the dense breast cases consisted of BI-RADS density of C and D. Each mammogram was labelled into normal and suspicious classifications. The normal mammogram was a mammogram with a BI-RADS classification of 1 or reported normal by the HPE result. On the other hand, the suspicious mammogram was a mammogram with BI-RADS classification of 2, 3, 4, 5, and 6 or reported as benign or malignant by the HPE result. Additionally, a mammogram with a BI-RADS classification of 0 was excluded from this study. Overall, there were 7,452 mammograms utilised in this study. About 1,651 mammograms were normal class and 5,801 mammograms were suspicious class.

3.2. Pre-processing steps

Each mammogram was pre-processed using a median filter, Otsu thresholding [23], and contrast limited adapted histogram equalisation (CLAHE). The median filter is a non-linear filtering method that aims to remove the noises in the image. Several studies had shown that the median filter had a good performance in mammograms in preserving the sharp edges and was robust to outliers [24–26]. Otsu thresholding is a type of clustering-based image thresholding which aims to binarise the image based on pixel intensities. The method had been shown to successfully remove the unwanted region with high intensities and the pectoral muscle in a mammogram, thus further improving mammogram classification and breast cancer detection [27,28]. Additionally, CLAHE was utilised to enhance the contrast of the mammogram. Several studies had proposed the use of this method as one of the pre-processing techniques to improve the predictive performance of breast cancer detection [29–31]. Lastly, the mammograms were rescaled, resized to 480x480, and their format was changed from DICOM to JPEG to reduce the size of the mammograms. Figure 1 illustrates the general flow of the pre-processing steps.

All the pre-processing steps were done in R version 4.2.1 [32]. *reticulate* [33] and *pydicom* [34] packages were used to read the mammogram into R. *nandb* [35], *EBImage* [36], and *autothresholdr* [37] packages were used to implement the median filter, CLAHE and resizing of the mammograms, and Otsu thresholding over the mammograms, respectively.

3.3. pre-trained network architecture

Thirteen pre-trained network architectures were applied based on the previous studies including MobileNets [38], MobileNetV2 [39], MobileNetV3Small [38], NasNetLarge [40], NasNetMobile [40], ResNet101 [41], ResNet101V2 [42], ResNet152 [41], ResNet152V2 [42], ResNet50 [41], ResNet50V2 [42], VGG16 [43], and VGG19 [43]. All pre-trained networks were run in R using *keras* [44] and *tensorflow* [45] packages.

The fine-tuning approach was used to customise the pre-trained network. The top layer with the largest parameters would be unfrozen layer by layer. The process would stop once the pre-trained network with the unfrozen current layer could not achieve a better performance compared to the pre-trained network with the unfrozen previous layer.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 March 2023 doi:10.20944/preprints202303.0015.v1

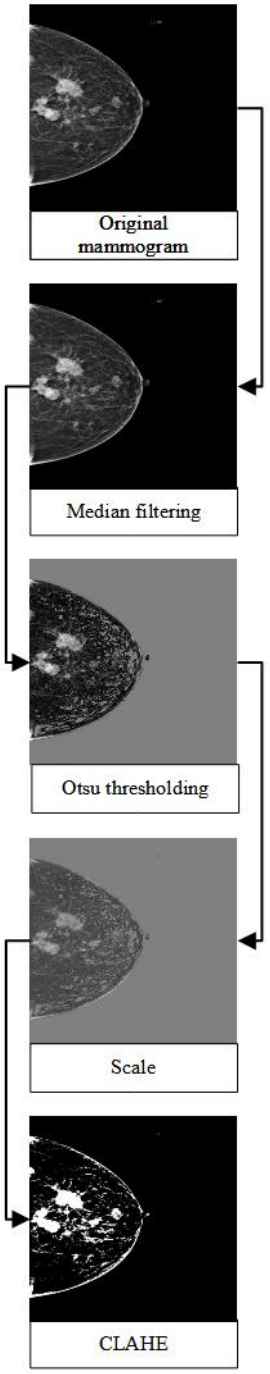

Figure 1. The general flow of the pre-processing techniques utilised in the study.

3.4. Model development and comparison

The data was split into three training-testing splits: (1) 70%-30%, (2) 80%-20%, and (3) 90%-10%. The validation dataset was set to 10% of each training dataset. Each mammogram was randomly classified into training, validation, and testing datasets. However, two stratification factors taken into consideration were the distribution of the breast density and mammogram classification. Thus, each training, validation, and testing dataset in each split was equally stratified and had an equal proportion of breast density (dense and non-dense) and mammogram classification (normal and suspicious).

Data augmentation and dropout were applied to overcome overfitting. Each mammogram was randomly flipped along a horizontal axis, rotated by a factor of 0.2 radians, and zoomed in or out by a factor of 0.05. The dropout was set at the rate of 0.5.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 March 2023 doi:10.20944/preprints202303.0015.v1

Additionally, class weight was used to overcome the class imbalance between normal and suspicious cases. The ratio of class weights used was 2.26 for normal and 0.64 for suspicious cases. Thus, the loss function heavily penalised the misclassification of the minority class (normal cases) compared to the misclassification of the majority class (suspicious cases). Binary cross-entropy was used as a loss function and Adam [46] algorithm was used as an optimiser. The learning rate was set to 1e-5. Lastly, a sigmoid activation function was used in the last layer to get the probability of the mammogram being suspicious. The network with the highest precision recall-area under the curve (PR-AUC) on the validation dataset was selected as the final model for each pre-trained network.

The evaluation criteria were applied to get the top fine-tuned pre-trained networks. The evaluation criteria utilised were the Youden J index > 0 and F1 score > 0.6. The candidates for the ensemble model were selected based on the PR-AUC, precision, and F1 score. Each ensemble model consisted of the top three pre-trained networks based on the three aforementioned performance metrics. The majority voting approach was utilised in each ensemble model to get the final prediction.

3.5. Performance metrics

Generally, the six performance metrics used in this study were PR-AUC, precision, F1 score, Youden J index, sensitivity, and specificity. The performance metrics were defined below:

$$Precision = \frac{TP}{TP + FP}$$

$$F1\ score = 2 \times \frac{precision \times recall}{precision + recall}$$

$$Youden\ J\ index = \ sensitivity + specificity - 1$$

$$Recall/sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

A true positive case was defined as a suspicious case and predicted suspicious by the network, while a true negative case was a normal case and predicted normal by the network. A false negative case was a suspicious case but predicted normal by the network, while a false positive case was a normal case but predicted suspicious by the network. All six performance metrics were aggregated across the three different splits and presented as mean and standard deviation (SD).

3.6. Performance across breast density

The final ensemble model was evaluated using the overall, dense, and non-dense testing datasets. The performance metrics were compared statistically using the Wilcoxon rank sum test. A p value < 0.05 indicated there was a significant difference in performance metrics between the dense and non-dense cases. Figure 2 illustrates the overall flow of the analysis in this study.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 March 2023 doi:10.20944/preprints202303.0015.v1

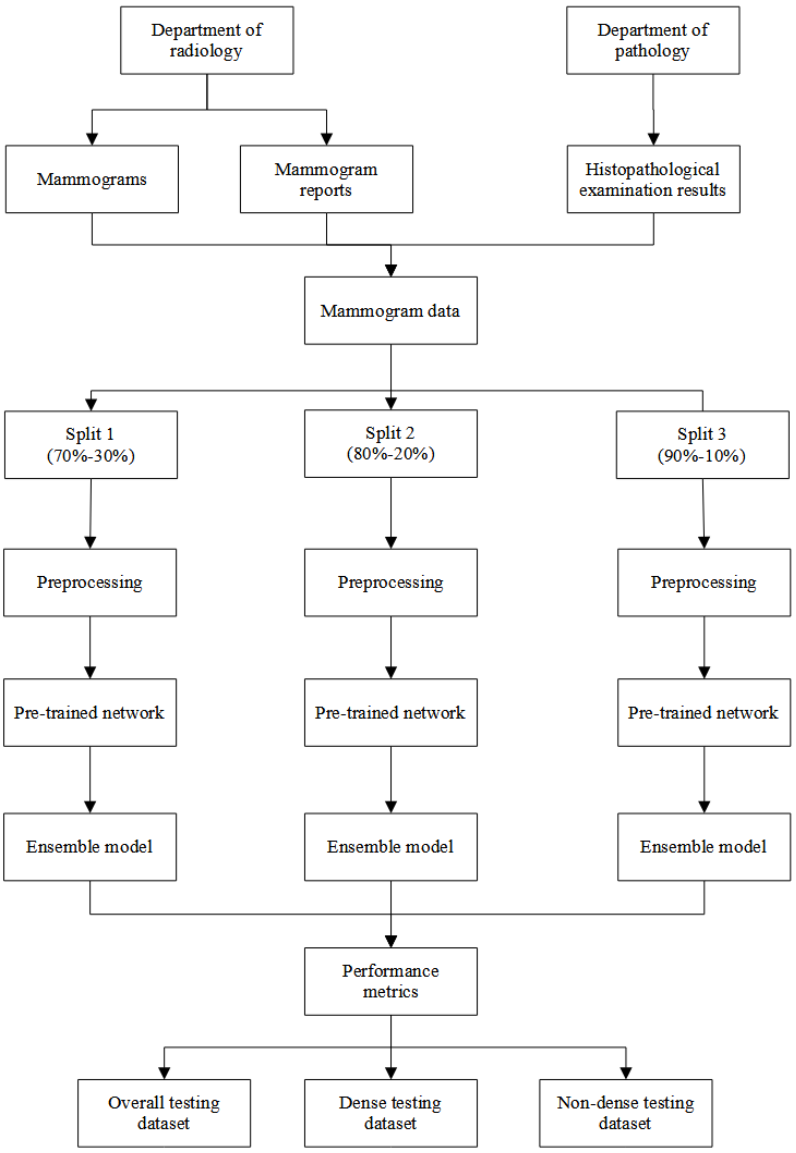

Figure 2. The overall flow of the analysis in this study.

4. Results

4.1. Model development

Thirteen pre-trained networks were developed and fine-tuned for breast abnormality detection in this study. Table 2 presents all the network architectures utilised in this study. The networks with the highest mean of PR-AUC, precision, F1 score, and Youden J index were ResNet101V2 and ResNet152, MobileNetV3Small and ResNet152, ResNet101, and ResNet152 and ResNet152V2, respectively. After the application of the evaluation criteria, only six networks remained out of thirteen pre-trained networks. Figure 3 presents all the selected pre-trained networks.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 March 2023 doi:10.20944/preprints202303.0015.v1

Table 2. Performance of fine-tuned pre-trained network for breast abnormality.

Architecture	PR-AUC (Mean, SD)	Precision (Mean, SD)	F1 score (Mean, SD)	Youden J index (Mean, SD)
MobileNets	0.79 (0.01)	0.79 (0.00)	0.49 (0.07)	0.02 (0.01)
MobileNetV2	0.79 (0.00)	0.79 (0.01)	0.46 (0.11)	0.02 (0.04)
MobileNetV3Small	0.80 (0.01)	0.81 (0.02)	0.56 (0.09)	0.06 (0.04)
NasNetLarge	0.80 (0.03)	0.80 (0.03)	0.68 (0.09)	0.06 (0.09)
NasNetMobile	0.79 (0.02)	0.79 (0.02)	0.67 (0.06)	0.03 (0.05)
ResNet101	0.80 (0.03)	0.79 (0.01)	0.73 (0.08)	0.04 (0.04)
ResNet101V2	0.81 (0.01)	0.79 (0.01)	0.61 (0.07)	0.02 (0.03)
ResNet152	0.81 (0.01)	0.81 (0.01)	0.65 (0.04)	0.07 (0.03)
ResNet152V2	0.80 (0.03)	0.80 (0.03)	0.60 (0.17)	0.07 (0.07)
ResNet50	0.80 (0.03)	0.78 (0.02)	0.66 (0.08)	0.01 (0.03)
ResNet50V2	0.80 (0.03)	0.80 (0.01)	0.67 (0.01)	0.05 (0.03)
VGG16	0.79 (0.03)	0.77 (0.04)	0.61 (0.14)	-0.01 (0.08)
VGG19	0.78 (0.02)	0.78 (0.01)	0.57 (0.11)	0.00 (0.04)

PR-AUC=precision recall-area under the curve
SD=standard deviation

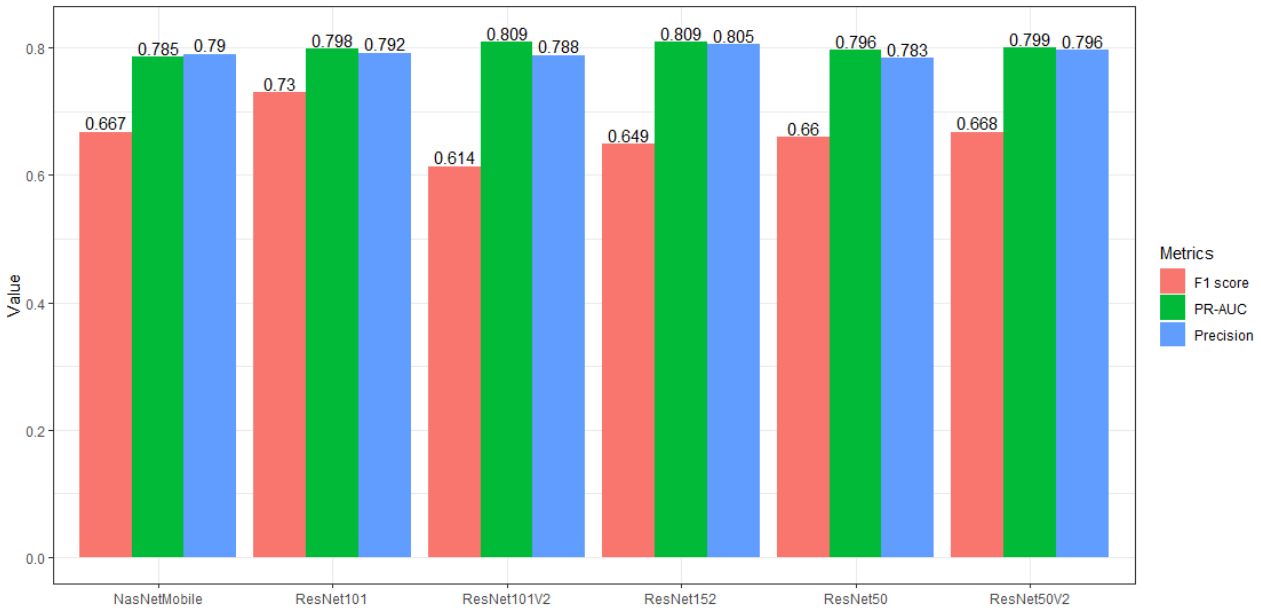

Figure 3. The top fine-tuned pre-trained network for breast abnormality detection.

4.2. Ensemble transfer learning

Three ensemble models were developed using a majority voting approach. Ensemble model 1 consisted of Resnet101, NasNetMobile, and ResNet50V2. Ensemble model 2 consisted of Resnet101V2, Resnet152, and ResNet50V2. Additionally, ensemble model 3 consisted of Resnet101, Resnet152, and ResNet50V2. Ensemble model 1, 2, and 3 was developed based on top F1 score, PR-AUC, and precision, respectively. Table 3 compares the performance metrics of the ensemble models and each candidate network. Ensemble model 3 had the highest mean precision and Youden J index, while ResNet101 had the highest mean F1 score. Thus, ensemble model 3 was selected as the final model in this study.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 March 2023 doi:10.20944/preprints202303.0015.v1

Table 3. Performance comparison between the ensemble transfer learning and the individual model for detection of breast abnormality.

Model	Precision (Mean, SD)	F1 score (Mean, SD)	Youden J index (Mean, SD)
Ensemble model 1	0.81 (0.01)	0.65 (0.01)	0.09 (0.03)
Ensemble model 2	0.81 (0.01)	0.66 (0.01)	0.09 (0.04)
Ensemble model 3	0.82 (0.01)	0.68 (0.01)	0.12 (0.03)
NasNetMobile	0.79 (0.02)	0.67 (0.06)	0.03 (0.05)
ResNet101	0.79 (0.01)	0.73 (0.08)	0.04 (0.04)
ResNet101V2	0.79 (0.01)	0.61 (0.07)	0.02 (0.03)
ResNet152	0.81 (0.01)	0.65 (0.04)	0.07 (0.03)
ResNet50V2	0.80 (0.01)	0.67 (0.01)	0.05 (0.03)

PR-AUC=precision recall-area under the curve
SD=standard deviation
Ensemble model 1 = Resnet101 + NasNetMobile + ResNet50V2
Ensemble model 2 = Resnet101V2 + Resnet152 + ResNet50V2
Ensemble model 3 = Resnet101 + Resnet152 + ResNet50V2

4.3. Performance across breast density

The final ensemble model was evaluated across the breast density. Table 4 presents the descriptive performance of the model, while Table 5 presents the result of the performance comparison of the model across dense and non-dense breast cases. The final model had slightly higher performance metrics in the dense breast cases compared to the non-dense breast cases. However, the result of the Wilcoxon rank sum test revealed that there was no significant difference between the dense and non-dense breasts across all performance metrics.

Table 4. The descriptive performance of the final ensemble model across breast density.

Metrics	Overall	Dense	Non-dense
Precision	0.82 (0.01)	0.86 (0.01)	0.77 (0.00)
F1 score	0.68 (0.01)	0.75 (0.01)	0.60 (0.02)
Youden J Index	0.12 (0.03)	0.21 (0.04)	0.03 (0.03)
Sensitivity	0.58 (0.02)	0.67 (0.01)	0.49 (0.03)
Specificity	0.54 (0.02)	0.54 (0.03)	0.54 (0.01)

Table 5. The performance comparison of the final ensemble model across breast density Wilcoxon rank sum test.

Metrics	Dense Median (IQR)	Non-dense Median (IQR)	W statistics	*P* value
Precision	0.86 (0.01)	0.77 (0.00)	9	0.1
F1 score	0.75 (0.01)	0.60 (0.02)	9	0.1
Youden J Index	0.22 (0.04)	0.03 (0.03)	9	0.1
Sensitivity	0.67 (0.01)	0.49 (0.03)	9	0.1
Specificity	0.55 (0.03)	0.54 (0.01)	6	0.7

IQR=interquartile range

5. Discussion

The final ensemble model in this study displayed a good performance with a precision of 0.82. Several studies achieved better precision metrics compared to this study ranging from 0.84 to 0.97 [11,12,19]. However, all these studies utilised publicly available datasets. Studies that used publicly available datasets had been shown to have a better performance compared to those that used primary datasets [47]. The data utilised in this study was mildly imbalanced. The proportion of minority class or normal mammograms was 22% of the total dataset. Thus, commonly used performance metrics such as accuracy and receiver operating curve-area under the curve (ROC-AUC) were not appropriate in

 doi:10.20944/preprints202303.0015.v1

this study. However, the data used in this study was collected from the department of radiology and pathology. Therefore, the performance presented in this study is more realistic and reflective of the actual performance of the deep learning model on mammographic data for breast abnormality detection. Notably, the performance of the final ensemble model was just slightly better than the initial fine-tuned pre-trained networks, especially compared to MobileNetV3Small and ResNet152 (Table 2 and Table 3) in this study. However, other studies that implemented the ensemble pre-trained network showed a better performance with an accuracy of 0.98 compared to each candidate network with an average accuracy of 0.94 [48]. This study utilised a microscopic image dataset to classify breast cancer.

This final ensemble model in this study also presented a balanced performance between specificity and sensitivity with an F1 score of 0.68. Theoretically, the relationship between the early two metrics is inversely proportionate [49]. A diagnostic tool with a high sensitivity typically had a low specificity, and vice versa. Thus, a balanced performance between the metrics was preferred, though any cut-off values are yet to be established. Further evaluation of the ensemble model across breast density revealed that there was no significant performance difference between dense and non-dense cases (Table 5). Breast density of mammographic density reduced the sensitivity of mammograms and increased the risk of breast [50,51]. Since Asian women tend to have more dense breasts compared to other ancestries [8], this factor plays a significant role in the screening and diagnosis of breast cancer in this population. The performance of any screening or diagnostic tool that utilised mammography should be evaluated in regard to breast density.

This study utilised mammographic data collected from a university-based hospital. The data were further evaluated by a radiologist and a pathologist. Thus, the data utilised in this study was good quality data and reflective of the actual cases in the hospital. Despite these strengths, this study suffered a mild imbalanced classification. Hence, common performance metrics such as accuracy and ROC-AUC were not appropriate to be used in this study. Consequently, the utilisation of different performance metrics made a comparison to other studies slightly challenging. Thus, future studies should try to achieve an imbalanced dataset. Moreover, future studies should include more hospitals, thus increasing the sample size of the study. Generally, a larger sample size may further improve the performance of the deep learning model.

5. Conclusions

This study explores the use of ensemble pre-trained networks or transfer learning for the purpose of breast abnormality detection. This model could be deployed as a supplementary diagnostic tool to radiologists, thus, reducing their workloads. Additionally, the use of the supplementary diagnostic tool in the medical workflow would improve the efficiency of breast cancer diagnosis, which in turn, accelerates the treatment and management for urgent cases. Furthermore, the use of this model may give radiologists more time to spend on the cases classified as suspicious rather than normal cases. Additionally, this model can be implemented in the medical workflow as a supplementary diagnostic tool for radiologists, thus further improves the efficiency in the management and diagnosis of the disease.

Author Contributions: Conceptualization, Tengku Muhammad Hanis, Nur Intan Raihana Ruhaiyem, Wan Nor Arifin and Kamarul Imran Musa; Data curation, Tengku Muhammad Hanis, Nur Intan Raihana Ruhaiyem and Wan Nor Arifin; Formal analysis, Tengku Muhammad Hanis, Nur Intan Raihana Ruhaiyem and Kamarul Imran Musa; Funding acquisition, Kamarul Imran Musa; Investigation, Juhara Haron and Wan Faiziah Wan Abdul Rahman; Methodology, Tengku Muhammad Hanis, Nur Intan Raihana Ruhaiyem and Wan Nor Arifin; Project administration, Kamarul Imran Musa; Resources, Juhara Haron and Wan Faiziah Wan Abdul Rahman; Supervision, Juhara Haron and Rosni Abdullah; Validation, Juhara Haron, Wan Faiziah Wan Abdul Rahman and Rosni Abdullah; Visualization, Tengku Muhammad Hanis and Kamarul Imran Musa; Writing – original draft, Tengku Muhammad Hanis; Writing – review & editing, Nur Intan Raihana Ruhaiyem, Wan Nor Arifin, Juhara Haron and Kamarul Imran Musa.

Funding: This research was funded by the Fundamental Research Grant Scheme (FRGS), Ministry of Higher Education, Malaysia (FRGS/1/2019/SKK03/USM/02/1).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the human research ethics committee of Universiti Sains Malaysia (JEPeM) on 19 November 2019 (USM/JEPeM/19090536).

Informed Consent Statement: Patient consent was waived due to the retrospective nature of this study and the use of secondary data.

Data Availability Statement: The data are available upon reasonable request to the corresponding author.

Acknowledgments: We thank all staff and workers in the Department of Radiology and Department of Pathology in Hospital Universiti Sains Malaysia for facilitating the data collection and extraction process.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Hosny, A.; Parmar, C.; Quackenbush, J.; Schwartz, L.H.; Aerts, H.J.W.L. Artificial intelligence in radiology. *Nat. Rev. Cancer* **2018**, *18*, 500–510, doi:10.1038/s41568-018-0016-5.
2. Seely, J.M.; Alhassan, T. Screening for Breast Cancer in 2018—What Should We be Doing Today? *Curr. Oncol.* **2018**, *25*, 115–124, doi:10.3747/co.25.3770.
3. Rodríguez-Ruiz, A.; Krupinski, E.; Mordang, J.-J.; Schilling, K.; Heywang-Köbrunner, S.H.; Sechopoulos, I.; Mann, R.M. Detection of Breast Cancer with Mammography: Effect of an Artificial Intelligence Support System. *Radiology* **2019**, *290*, 305–314, doi:10.1148/radiol.2018181371.
4. Chollet, F.; Kalinowski, T.; Allaire, J.J. *Deep learning with R*; 2nd ed.; Manning: Shelter, 2022;
5. Iman, M.; Rasheed, K.; Arabnia, H.R. A Review of Deep Transfer Learning and Recent Advancements. *arXiv* **2022**.
6. Ayana, G.; Dese, K.; Choe, S. Transfer Learning in Breast Cancer Diagnoses via Ultrasound Imaging. *Cancers (Basel).* **2021**, *13*, 738, doi:10.3390/cancers13040738.
7. Hanis, T.M.; Arifin, W.N.; Haron, J.; Wan Abdul Rahman, W.F.; Ruhaiyem, N.I.R.; Abdullah, R.; Musa, K.I. Factors Influencing Mammographic Density in Asian Women: A Retrospective Cohort Study in the Northeast Region of Peninsular Malaysia. *Diagnostics* **2022**, *12*, 860, doi:10.3390/diagnostics12040860.
8. Nazari, S.S.; Mukherjee, P. An overview of mammographic density and its association with breast cancer. *Breast Cancer* **2018**, *25*, 259–267, doi:10.1007/s12282-018-0857-5.
9. Weigel, S.; Heindel, W.; Heidrich, J.; Hense, H.-W.; Heidinger, O. Digital mammography screening: sensitivity of the programme dependent on breast density. *Eur. Radiol.* **2017**, *27*, 2744–2751, doi:10.1007/s00330-016-4636-4.
10. Fiorica, J. V Breast Cancer Screening , Mammography , and Other Modalities. *Clin. Obstet. Gynecol. Gynecol.* **2016**, *59*, 688–709.
11. Saber, A.; Sakr, M.; Abo-Seida, O.M.; Keshk, A.; Chen, H. A Novel Deep-Learning Model for Automatic Detection and Classification of Breast Cancer Using the Transfer-Learning Technique. *IEEE Access* **2021**, *9*, 71194–71209, doi:10.1109/ACCESS.2021.3079204.
12. Khamparia, A.; Bharati, S.; Podder, P.; Gupta, D.; Khanna, A.; Phung, T.K.; Thanh, D.N.H. Diagnosis of breast cancer based on modern mammography using hybrid transfer learning. *Multidimens. Syst. Signal Process.* **2021**, *32*, 747–765, doi:10.1007/s11045-020-00756-7.
13. Guan, S.; Loew, M. Using generative adversarial networks and transfer learning for breast cancer detection by convolutional neural networks. In Proceedings of the Medical Imaging 2019: Imaging Informatics for Healthcare, Research, and Applications; 2019; Vol. 10954, p. 48.
14. Mendel, K.; Li, H.; Sheth, D.; Giger, M. Transfer Learning From Convolutional Neural Networks for Computer-Aided

Diagnosis: A Comparison of Digital Breast Tomosynthesis and Full-Field Digital Mammography. *Acad. Radiol.* **2019**, *26*, 735–743, doi:10.1016/j.acra.2018.06.019.

15. Falconi, L.G.; Perez, M.; Aguilar, W.G.; Conci, A. Transfer learning and fine tuning in breast mammogram abnormalities classification on CBIS-DDSM database. *Adv. Sci. Technol. Eng. Syst.* **2020**, *5*, 154–165, doi:10.25046/aj050220.
16. Guan, S.; Loew, M. Breast Cancer Detection Using Transfer Learning in Convolutional Neural Networks. In Proceedings of the 2017 IEEE Applied Imagery Pattern Recognition Workshop (AIPR); 2017; pp. 1–8.
17. Yu, X.; Wang, S.-H. Abnormality Diagnosis in Mammograms by Transfer Learning Based on ResNet18. *Fundam. Informaticae* **2019**, *168*, 219–230, doi:10.3233/FI-2019-1829.
18. Falconi, L.G.; Perez, M.; Aguilar, W.G. Transfer Learning in Breast Mammogram Abnormalities Classification with Mobilenet and Nasnet. *Int. Conf. Syst. Signals, Image Process.* **2019**, 109–114, doi:10.1109/IWSSIP.2019.8787295.
19. Ansar, W.; Shahid, A.R.; Raza, B.; Dar, A.H. Breast Cancer Detection and Localization Using MobileNet Based Transfer Learning for Mammograms. In *International Symposium on …*; Springer, 2020; pp. 11–21.
20. Jiang, F.; Liu, H.; Yu, S.; Xie, Y. Breast mass lesion classification in mammograms by transfer learning. In Proceedings of the 5th International Conference on Bioinformatics and Computational Biology, ICBCB 2017; Association for Computing Machinery: Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China, 2017; pp. 59–62.
21. Mednikov, Y.; Nehemia, S.; Zheng, B.; Benzaquen, O.; Lederman, D. Transfer Representation Learning using Inception-V3 for the Detection of Masses in Mammography. In Proceedings of the 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 2018; Vol. 2018-July, pp. 2587–2590.
22. Pattanaik, R.K.; Mishra, S.; Siddique, M.; Gopikrishna, T.; Satapathy, S. Breast Cancer Classification from Mammogram Images Using Extreme Learning Machine-Based DenseNet121 Model. *Genet. Res. (Camb).* **2022**, *2022*, doi:10.1155/2022/2731364.
23. Otsu, N. A Threshold Selection Method from Gray-Level Histograms. *IEEE Trans. Syst. Man. Cybern.* **1979**, *9*, 62–66, doi:10.1109/TSMC.1979.4310076.
24. Kshema; George, M.J.; Dhas, D.A.S. Preprocessing filters for mammogram images: A review. In Proceedings of the 2017 Conference on Emerging Devices and Smart Systems (ICEDSS); 2017; pp. 1–7.
25. George, M.J.; Sankar, S.P. Efficient preprocessing filters and mass segmentation techniques for mammogram images. In Proceedings of the 2017 IEEE International Conference on Circuits and Systems (ICCS); 2017; Vol. 2018-Janua, pp. 408–413.
26. Lu, H.-C.; Loh, E.-W.; Huang, S.-C. The Classification of Mammogram Using Convolutional Neural Network with Specific Image Preprocessing for Breast Cancer Detection. In Proceedings of the 2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD); 2019; pp. 9–12.
27. Omer, A.M.; Elfadil, M. Preprocessing of Digital Mammogram Image Based on Otsu's Threshold. *Am. Sci. Res. J. Eng. Technol. Sci.* **2017**, *37*, 220–229.
28. Khairnar, S.; Thepade, S.D.; Gite, S. Effect of image binarization thresholds on breast cancer identification in mammography images using OTSU, Niblack, Burnsen, Thepade's SBTC. *Intell. Syst. with Appl.* **2021**, *10–11*, 200046, doi:10.1016/j.iswa.2021.200046.
29. Lbachir, I.A.; Es-Salhi, R.; Daoudi, I.; Tallal, S. A New Mammogram Preprocessing Method for Computer-Aided Diagnosis Systems. In Proceedings of the 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA); 2017; Vol. 2017-Octob, pp. 166–171.
30. Mat Radzi, S.F.; Abdul Karim, M.K.; Saripan, M.I.; Abd Rahman, M.A.; Osman, N.H.; Dalah, E.Z.; Mohd Noor, N. Impact of Image Contrast Enhancement on Stability of Radiomics Feature Quantification on a 2D Mammogram Radiograph. *IEEE Access* **2020**, *8*, 127720–127731, doi:10.1109/ACCESS.2020.3008927.
31. Kharel, N.; Alsadoon, A.; Prasad, P.W.C.; Elchouemi, A. Early diagnosis of breast cancer using contrast limited adaptive histogram equalization (CLAHE) and Morphology methods. In Proceedings of the 2017 8th International Conference on Information and Communication Systems (ICICS); 2017; pp. 120–124.

32. R Core Team R: A Language and Environment for Statistical Computing 2022.
33. Ushey, K.; Allaire, J.J.; Tang, Y. reticulate: Interface to "Python" 2023.
34. Mason, D. SU-E-T-33: Pydicom: An Open Source DICOM Library. *Med. Phys.* **2011**, *38*, 3493–3493, doi:10.1118/1.3611983.
35. Nolan, R.; Alvarez, L.A.J.; Elegheert, J.; Iliopoulou, M.; Jakobsdottir, G.M.; Rodriguez-Muñoz, M.; Aricescu, A.R.; Padilla-Parra, S. nandb—number and brightness in R with a novel automatic detrending algorithm. *Bioinformatics* **2017**, *33*, 3508–3510, doi:10.1093/bioinformatics/btx434.
36. Pau, G.; Fuchs, F.; Sklyar, O.; Boutros, M.; Huber, W. EBImage—an R package for image processing with applications to cellular phenotypes. *Bioinformatics* **2010**, *26*, 979–981, doi:10.1093/bioinformatics/btq046.
37. Landini, G.; Randell, D.A.; Fouad, S.; Galton, A. Automatic thresholding from the gradients of region boundaries. *J. Microsc.* **2017**, *265*, 185–195, doi:10.1111/jmi.12474.
38. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**.
39. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018; pp. 4510–4520.
40. Zoph, B.; Vasudevan, V.; Shlens, J.; Le, Q. V. Learning Transferable Architectures for Scalable Image Recognition. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018; pp. 8697–8710.
41. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016; pp. 770–778.
42. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity Mappings in Deep Residual Networks. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer International Publishing, 2016; Vol. 9908, pp. 630–645 ISBN 9783319464923.
43. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*; Yoshua, B., LeCun, Y., Eds.; San Diego, CA, 2015; pp. 1–14.
44. Allaire, J.J.; Chollet, F. keras: R Interface to "Keras" 2022.
45. Allaire, J.J.; Tang, Y. tensorflow: R Interface to "TensorFlow" 2022.
46. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings; 2014; pp. 1–15.
47. Hanis, T.M.; Islam, M.A.; Musa, K.I. Diagnostic Accuracy of Machine Learning Models on Mammography in Breast Cancer Classification: A Meta-Analysis. *Diagnostics* **2022**, *12*, 1643, doi:10.3390/diagnostics12071643.
48. Khan, S.; Islam, N.; Jan, Z.; Ud Din, I.; Rodrigues, J.J.P.C. A novel deep learning based framework for the detection and classification of breast cancer using transfer learning. *Pattern Recognit. Lett.* **2019**, *125*, 1–6, doi:10.1016/j.patrec.2019.03.022.
49. Shreffler, J.; Huecker, M.R. *Diagnostic Testing Accuracy: Sensitivity, Specificity, Predictive Values and Likelihood Ratios*; StatPearls Publishing, Treasure Island (FL), 2022;
50. Lynge, E.; Vejborg, I.; Andersen, Z.; von Euler-Chelpin, M.; Napolitano, G. Mammographic Density and Screening Sensitivity, Breast Cancer Incidence and Associated Risk Factors in Danish Breast Cancer Screening. *J. Clin. Med.* **2019**, *8*, 2021, doi:10.3390/jcm8112021.
51. Sherratt, M.J.; McConnell, J.C.; Streuli, C.H. Raised mammographic density: Causative mechanisms and biological consequences. *Breast Cancer Res.* **2016**, *18*, 1–9, doi:10.1186/S13058-016-0701-9.