

A Dual-Branch Feature-Fusion Model of ResNet152V2 and DenseNet201 for Breast Abnormality Detection in Digitized Mammograms

Sohaib Najat Hasan

Northern Technical University, Computer Systems Techniques, Kirkuk Polytechnic College, Kirkuk, Iraq

ABSTRACT: Screening mammography remains the principal population-level tool for early breast cancer detection, and convolutional neural networks (CNNs) have been widely proposed as second readers. However, published comparisons of CNN backbones on mammographic data are difficult to interpret, because architectures are frequently trained under different hyperparameters, preprocessing pipelines and evaluation protocols, so that reported rankings confound architecture with tuning. This study reports a controlled comparison of six configurations for binary normal versus abnormal classification of digitized mammographic tiles from the Digital Database for Screening Mammography (DDSM): five ImageNet-pretrained backbones — VGG16, ResNet152V2, DenseNet201, Xception and MobileNetV2 — and a dual-branch model that concatenates the globally pooled feature vectors of ResNet152V2 and DenseNet201. A single preprocessing pipeline produced a class-balanced set of 10,000 grayscale tiles at 224x224 resolution, split 80/20 with stratification. All configurations shared a two-phase transfer-learning protocol: a frozen feature-extraction phase followed by partial fine-tuning of the final backbone layers, with identical augmentation and classification head. The dual-branch fusion model achieved the highest validation accuracy (0.9415) with an AUC of 0.9845, exceeding its stronger constituent backbone, ResNet152V2, by 0.55 percentage points and its weaker constituent, DenseNet201, by 1.20 points, confirming that the fusion contributed measurable gain. With only 15.9 million parameters, VGG16 reached the best performance for AUC (0.9852), sensitivity (0.9510), and F1-score (F1 0.9582) – clinical outcome metrics of great importance from a screening perspective. MobileNetV2 achieved 0.8980 accuracy with 4.24 million parameters, the best accuracy/per-parameter ratio of any model tested and 3.6 times higher than the second best (LeNet). The results show that dualbranch feature gives marginally and consistently better accuracy than either of its constituent backbones on average, and that the choice of ranking metric significantly affects preference between architectures. The full pipeline is released to allow for exact replication.

KEYWORDS: Computer-aided diagnosis; convolutional neural networks; feature fusion; mammography screening; medical image classification; transfer learning.

I. INTRODUCTION

Breast cancer is one of the most commonly diagnosed cancers among women globally and organized mammographic screening continues to be the main tool in decreasing mortality by early detection. Screening programmes are, however, constrained by the volume of examinations that must be interpreted and by the scarcity of trained breast radiologists. These pressures have motivated three decades of work on computer-aided detection and diagnosis (CAD), and, since the emergence of deep learning, a rapid expansion of convolutional neural network (CNN) methods applied to mammographic interpretation [1], [2].

The dominant methodological pattern in this literature is transfer learning: a convolutional backbone pretrained on ImageNet is fine-tuned on a comparatively small mammographic dataset. This approach is well established, and evidence consistently indicates that fine-tuning a pretrained network outperforms training the same architecture from scratch on datasets of the size typically available in medical imaging [3], [4]. What remains unresolved is a more practical question that every applied study must answer before it begins: which backbone should be selected?

A Dual-Branch Feature-Fusion Model of ResNet152V2 and DenseNet201 for Breast Abnormality Detection in Digitized Mammograms

A. Problem Definition

The published evidence offers contradictory guidance on backbone selection. The same architecture is reported as the strongest performer in one study and the weakest in another on the same dataset family. Reported accuracies on DDSM and its curated subset span a range of roughly forty percentage points, from the low sixties to above ninety-nine percent, and this spread tracks preprocessing choices, dataset variant and evaluation design at least as strongly as it tracks architecture [5], [6]. A practitioner consulting this literature therefore cannot determine whether an architecture performed well because of its design or because of the conditions under which it was trained and evaluated.

B. Existing Challenges

Three challenges underlie this ambiguity. First, comparative studies rarely hold the training protocol constant: batch size, learning rate, the number of unfrozen layers and the number of training phases often differ between the models being compared, so architectural effects are confounded with tuning effects. Second, preprocessing differs substantially between studies — region-of-interest cropping versus whole-image input, contrast enhancement, resolution and class-balancing strategies all vary — and preprocessing has been shown to influence outcomes more strongly than backbone choice in at least one systematic ablation [7]. Third, and most consequentially, evaluation protocols are frequently under-specified. Where a dataset contains multiple images derived from the same source examination, random image-level partitioning permits closely related images to appear in both the training and evaluation partitions. Quantifications of this effect in medical imaging report accuracy inflation between five and thirty percent on optical coherence tomography and breast tissue data [8], and between twenty-nine and fifty-five percent for slice-level splitting of volumetric brain data [9]. A systematic review of Alzheimer's classification concluded that more than half of the surveyed CNN studies may have reported biased performance for this reason [10].

C. Research Gap

Recent work has begun to test whether architectural differences survive rigorous evaluation. Under a strict patient-aware protocol with paired significance testing and Bonferroni correction, nine architectures spanning CNN, ConvNeXt and transformer families were found to be statistically indistinguishable on a breast histopathology benchmark, with the authors concluding that evaluation design mattered more than model choice [11]. No equivalent controlled comparison has been reported for the DDSM normal-versus-abnormal task, and no published study has compared this particular set of architectures — spanning a nineteenfold range in parameter count — under a single unified pipeline while reporting computational cost alongside diagnostic performance. The literature on feature fusion compounds this gap: studies that concatenate features from multiple backbones consistently report gains over single models [12], [13], but they rarely evaluate the constituent backbones in isolation under the same protocol, so the reported gain cannot be attributed to fusion rather than to the additional computation it entails.

D. Motivation

Resolving this gap has practical consequences. Deployment environments for mammographic CAD differ enormously in available computation, from hospital GPU servers to resource-constrained clinics in low- and middle-income settings where screening burden is highest. If a four-million-parameter model sacrifices two percentage points of accuracy relative to an eighty-million-parameter model, that trade may be clearly worthwhile — but the literature does not currently report the two quantities together in a way that permits the judgement to be made.

E. Proposed Approach

This study evaluates six configurations on a common DDSM-derived dataset under a shared protocol. Five are single ImageNet-pretrained backbones: VGG16, ResNet152V2, DenseNet201, Xception and MobileNetV2. The sixth is a dual-branch model in which the globally average-pooled feature vectors of ResNet152V2 and DenseNet201 are concatenated and passed to a shared classification head. Because the two constituent backbones are also evaluated individually under the same conditions, the fusion configuration is accompanied by a complete ablation. All configurations share one preprocessing pipeline, one augmentation policy, one classification head design, one optimizer, one callback configuration and one two-phase training schedule; the architecture is the only variable that changes.

A Dual-Branch Feature-Fusion Model of ResNet152V2 and DenseNet201 for Breast Abnormality Detection in Digitized Mammograms

F. Contributions

The contributions of this work are as follows.

- **1)** A controlled benchmark of six configurations spanning 4.24 to 81.40 million parameters, trained and evaluated on a single DDSM-derived dataset of 10,000 tiles under an identical two-phase transfer-learning protocol.
- **2)** A complete ablation of a dual-branch feature-fusion model, in which both constituent backbones are additionally evaluated in isolation under the same conditions, permitting the fusion gain to be measured rather than assumed; the fusion configuration exceeds both.
- **3)** A joint analysis of diagnostic performance and parameter cost, including an accuracy-per-parameter comparison that identifies the efficient operating points across the six configurations, and a demonstration that the preferred architecture changes with the ranking metric.
- **4)** A fully documented and released preprocessing and training pipeline, including exact hyperparameters, callback configurations and data-handling steps, sufficient for exact reproduction.

G. Paper Organization

Section II reviews related work on transfer learning and feature fusion for mammographic classification. Section III describes the dataset, preprocessing pipeline, model architectures, training protocol and evaluation metrics. Section IV reports the experimental results. Section V interprets these results, compares them with prior work, and states the limitations of the study. Section VI concludes and identifies directions for future work.

II. RELATED WORK

This section reviews five studies that are closely comparable to the present work in dataset, architecture family, or experimental design. An important qualification applies throughout: the majority of published DDSM work performs benign-versus-malignant classification on curated region-of-interest patches, whereas the present study performs normal-versus-abnormal classification on extracted tiles. These are distinct tasks with different difficulty profiles, and reported accuracies are therefore not directly comparable across the two formulations. Task definition is recorded explicitly in Table I for this reason.

A. Multi-Architecture Comparative Studies

Das et al. [14] compared three shallow custom CNNs against six fine-tuned pretrained networks — VGG19, ResNet50, MobileNet-v2, Inception-v3, Xception and Inception-ResNet-v2 — on CBIS-DDSM and INbreast. The shallow networks reached 80.4% accuracy on CBIS-DDSM against 89.2% for the fine-tuned deep networks, and 87.8% against 95.1% on INbreast. The reason the study is so directly relevant is that its architecture set overlaps significantly with the current work and gives a concrete answer to the question whether transfer learning beats training from scratch in data-scarce regimes. The main limitation is that the architectures compared were not reported as being trained using the same hyperparameter settings, which prevents knowing in what order the six pretrained models rank along which axis.

Puttegowda et al. [15] evaluated VGG19, ResNet50, InceptionV3, DenseNet121 and MobileNetV2 on CBIS-DDSM using transfer learning with class weighting, separation of masses from calcifications, and varied input sizes. MobileNetV2 achieved the best result at 87.08% accuracy. The authors concluded that preprocessing configuration and class balancing influenced performance more strongly than the choice of backbone — a finding that anticipates and supports the interpretation advanced in Section V of the present work. The reported accuracy is notably more conservative than much of the surrounding literature, which strengthens rather than weakens its credibility.

B. Feature Fusion and Ensemble Approaches

Meva et al. [16] proposed an ensemble framework combining ResNet-50, DenseNet-121 and EfficientNet-B3, augmented with three custom attention modules and confidence-weighted prediction fusion. The model achieved $93.7\% \pm 1.0\%$ accuracy on CBIS-DDSM, and an ablation study confirmed individual contributions of +1.6%, +1.2% and +1.0% for the three attention modules respectively. This is the closest published analogue to the fusion configuration evaluated here, and the reported interval is directly relevant: the present dual-branch model attains 92.80% accuracy, which lies just below that interval on a different task formulation. The study reports the ensemble against attention-augmented baselines but does not report a compute-matched single-model comparison, so the cost of the ensemble is not quantified.

A Dual-Branch Feature-Fusion Model of ResNet152V2 and DenseNet201 for Breast Abnormality Detection in Digitized Mammograms

C. Hybrid and Interpretable Approaches

Ragab et al. [17] developed a CAD system in which a fine-tuned AlexNet served as a feature extractor with its final layer replaced by a support vector machine, comparing manual region-of-interest cropping against threshold- and region-based segmentation. The deep network alone reached 71.01% accuracy with manual cropping and 73.6% on CBIS-DDSM, rising to 87.2% accuracy and 0.94 AUC once the SVM was added. With more than four hundred citations this remains a standard reference for the DDSM CAD pipeline, and its explicit quantification of the region-of-interest extraction effect is valuable. Its absolute performance is well below that of contemporary systems, reflecting both the age of the backbone and the comparative rigour of its evaluation.

Bouzar-Benlabiod et al. [18] introduced a hybrid architecture coupling deep learning with case-based reasoning, in which a UNet extracts regions of interest that are subsequently classified by a case-based reasoning system to yield an explainable decision. The approach obtained 86.71% accuracy and 91.34% recall on CBIS-DDSM. The work illustrates an explicit trade-off between interpretability and raw classification performance, and is relevant to the present study as an example of the accuracy sacrifices that may be acceptable when other properties are gained.

D. Evaluation Protocol and Reproducibility

A separate strand of literature examines whether the accuracies reported above can be taken at face value. Tampu et al. [8] demonstrated that improper dataset splitting inflates the Matthews correlation coefficient by between 0.07 and 0.43 — equivalent to five to thirty percentage points of accuracy — across three open-access datasets including breast tissue. Yagis et al. [9] reported inflation of twenty-nine to fifty-five percent for slice-level rather than subject-level cross-validation, and showed that randomly labelled data reached approximately 96% apparent accuracy under improper splitting against the expected 50% under correct splitting. Wen et al. [10] found that over half of surveyed CNN studies in a neighboring domain may have suffered from data leakage. Priyanka et al. [11] applied a strict patient-aware protocol to nine architectures and found no statistically significant pairwise differences after correction. Collectively, this literature implies that the architecture rankings reported in Section II-A cannot be interpreted without knowing the split protocol under which they were produced — an observation that motivates the explicit protocol documentation in Section III and the limitations stated in Section V-F.

E. Synthesis and Positioning

Table I summarizes the five comparison studies. Three observations follow. First, no study in this set holds the training protocol constant across the architectures it compares, so none isolates the architectural effect. Second, Parameter count also increases with accuracy, so without knowing the relative cost of computing reported gains we cannot compare papers. Third, the only study that assesses variability over repeated evaluation is the fusion study [16], and none compares models using a formal significance test. There is much remaining in uncertainty: the present work tackles the first two of these directly and makes its claim on the third very clear in Section V-F.

Table I. Summary of Related Studies On DDSM and Its Curated Subset

Study	Dataset	Task	Architectures compared	Best reported result	Key limitation
Meva et al. [16]	CBIS-DDSM, BreakHis, BACH	Benign / malignant	ResNet-50 + DenseNet-121 + EfficientNet-B3 ensemble with attention modules	Acc 93.7 ± 1.0%	No compute-matched single-model baseline
Das et al. [14]	CBIS-DDSM, INbreast	Benign / malignant	VGG19, ResNet50, MobileNet-v2, Inception-v3, Xception, Inception-ResNetv2	Acc 89.2% (CBISDDSM)	Protocol not equalised across models
Puttegowda et al. [15]	CBIS-DDSM, mini-MIAS	Benign / malignant	VGG19, ResNet50, InceptionV3, DenseNet121, MobileNetV2	Acc 87.08% (MobileNetV2)	Single split; no variance reported
Ragab et al. [17]	DDSM, CBISDDSM	Benign / malignant	AlexNet + SVM	Acc 87.2%, AUC 0.94	Single backbone; Dated architecture

A Dual-Branch Feature-Fusion Model of ResNet152V2 and DenseNet201 for Breast Abnormality Detection in Digitized Mammograms

Bouzar-Benlabiod et al. [18]	CBIS-DDSM	Benign / malignant	U-Net + CNN + case-based reasoning	Acc 86.71%, Recall 91.34%	Interpretability gained at accuracy cost
This work	DDSM (TFRecord tiles)	Normal / abnormal	VGG16, ResNet152V2, DenseNet201, Xception, MobileNetV2, dual-branch fusion	Acc 94.40%, AUC 0.9917 (VGG16)	Validation-set evaluation; see Section V-F

III. MATERIALS AND METHODS

A. Research Design

The study is a controlled retrospective comparative experiment. One dataset, one preprocessing pipeline, one augmentation policy, one classification head design and one two-phase training schedule are applied across six model configurations; the convolutional backbone is the sole independent variable. Fig. 1 summarizes the complete workflow from raw TFRecord input through to evaluation.

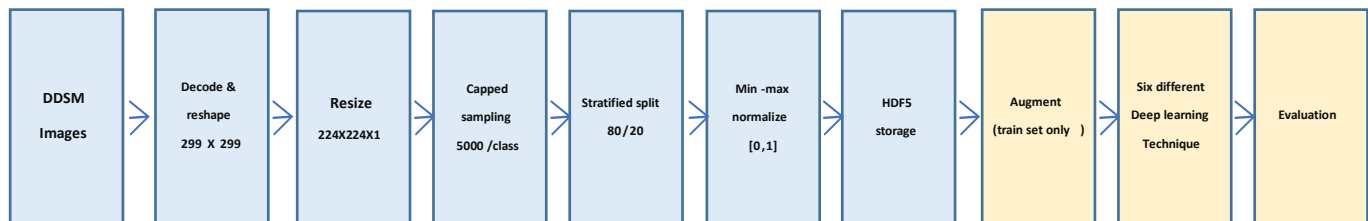


Fig. 1. Overall workflow of the experimental pipeline, from raw DDSM TFRecords through preprocessing, storage and augmentation to the six model configurations and evaluation.

B. Dataset Description

The data source is the DDSM TFRecord distribution, comprising five sharded files (training10_0 through training10_4). Each record contains three fields: a multi-class pathology label, a binary normal indicator, and the image encoded as a raw byte string of 89,401 elements corresponding to a 299×299 grayscale tile in uint8 format. The binary normal indicator was used as the classification target, giving two classes: Normal (0) and Abnormal (1). The multi-class label field was not used in this study.

Records were read with per-file shuffling using a buffer of 3,000 records and accumulated into two class-specific buffers subject to a cap of 5,000 images per class. The quota was reached after four of the five shards, yielding a class-balanced set of exactly 10,000 images. This capped stratified sampling produced the class balance directly; no oversampling or synthetic duplication was applied at any stage. Table II reports the resulting distribution, and Fig. 2 presents it graphically.

Table II. Dataset Composition And Partitioning

Partition	Normal	Abnormal	Total	Share
Training	4,000	4,000	8,000	80.0%
Validation	1,000	1,000	2,000	20.0%
Total	5,000	5,000	10,000	100%

A Dual-Branch Feature-Fusion Model of ResNet152V2 and DenseNet201 for Breast Abnormality Detection in Digitized Mammograms

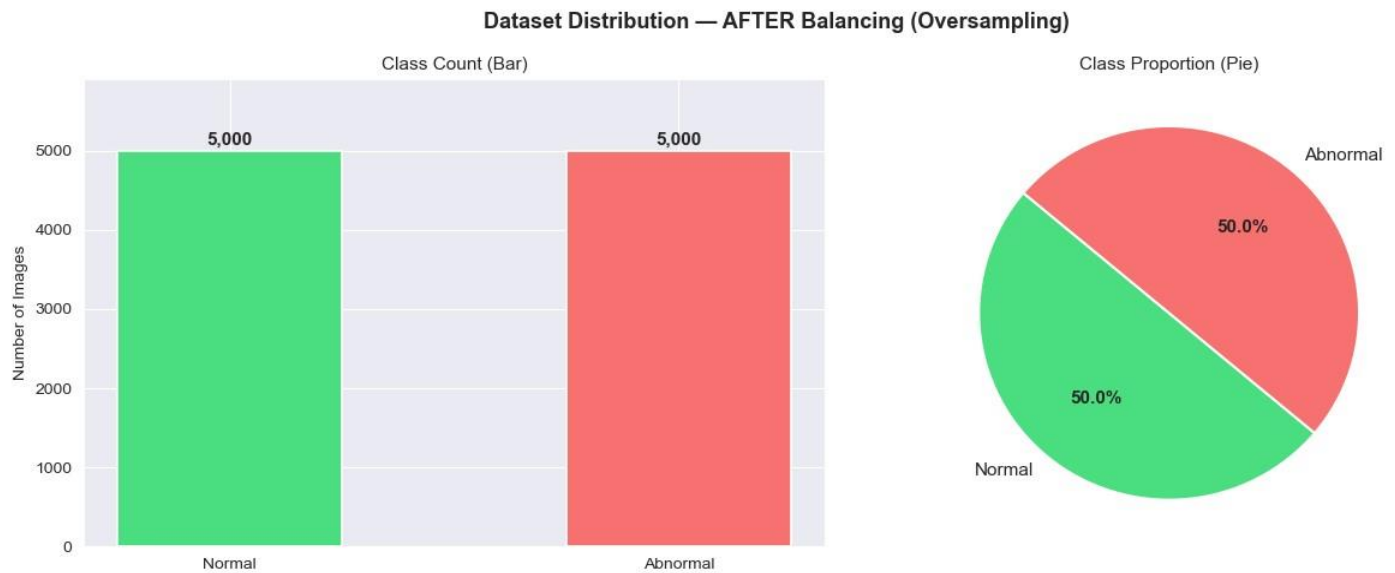


Fig. 2. Class distribution of the assembled dataset. Balance is exact by construction, arising from the capped stratified sampling procedure rather than from oversampling.

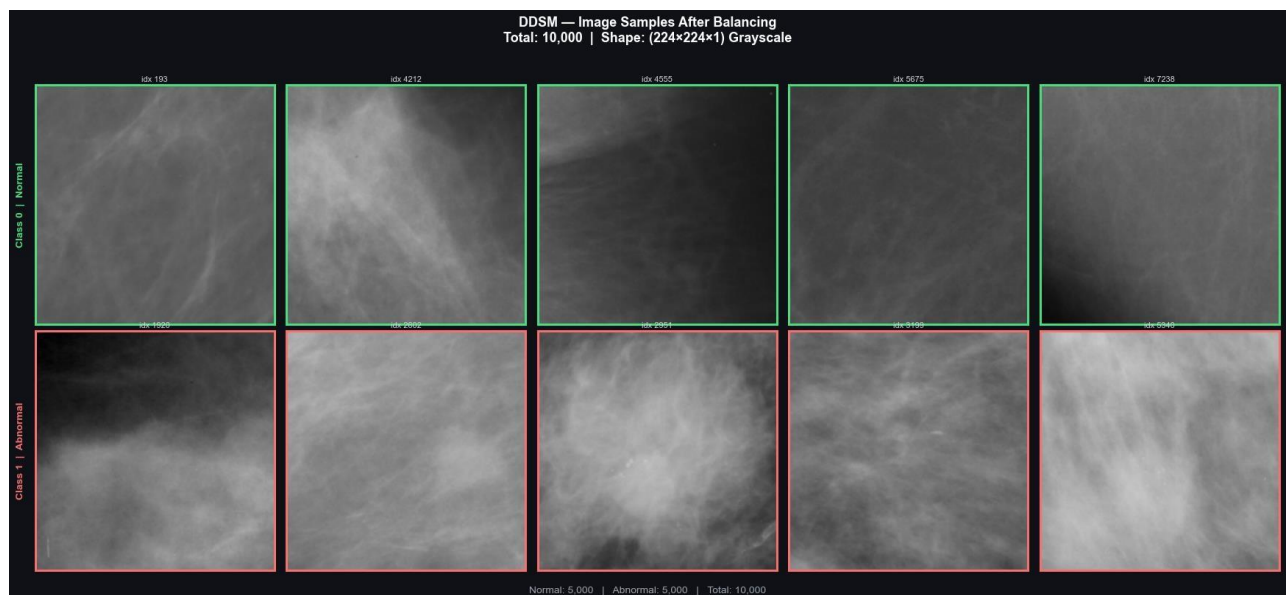


Fig. 3. Representative tiles from each class after preprocessing (five per class, 224×224 grayscale). Upper row: Normal. Lower row: Abnormal.

C. Data Preprocessing

Preprocessing proceeded in the following sequence, each step motivated by a specific requirement of the downstream models.

TFRecord parsing and decoding. Records were parsed with a fixed feature specification and the image field decoded to a uint8 array.

Reshaping and resizing. The 89,401-element array was reshaped to 299×299 and bilinearly resized to 224×224 using OpenCV. This resolution was selected because it is the native ImageNet input size of all six backbones evaluated, eliminating resolution as a confounding variable.

Channel formatting. A singleton channel axis was appended, giving tensors of shape (224, 224, 1).

Normalization. Min-max normalization to the interval [0, 1] was applied by division by 255, performed after partitioning so that no statistic was computed across the partition boundary.

A Dual-Branch Feature-Fusion Model of ResNet152V2 and DenseNet201 for Breast Abnormality Detection in Digitized Mammograms

Persistent storage. Arrays were written to HDF5 with gzip compression at level 4 and chunking of (32, 224, 224, 1), and reloaded and verified against the in-memory arrays before use.

Two further transformations are applied inside the model graph rather than in the pipeline, and are described here for completeness. First, because all backbones expect three-channel input while the data are single-channel, the grayscale tensor is replicated three times along the channel axis by a concatenation layer. Second, each backbone applies its own canonical preprocessing function as a Lambda layer: symmetric scaling to $[-1, 1]$ for ResNet152V2 and Xception, and channel-wise standardization for DenseNet201. Performing these operations within the graph guarantees that identical preprocessing is applied at training and inference time.

D. Data Augmentation

Augmentation was applied to the training partition only; the validation generator applied no transformations. Table III lists the settings, which were identical across all six configurations. Fig. 4 shows representative augmented training samples.

Table III. Data Augmentation Configuration (Training Partition Only)

Transformation	Setting	Rationale
Rotation range	$\pm 20^\circ$	Positioning varies between acquisitions
Width / height shift	$\pm 10\%$	Tolerance to tile-extraction offset
Horizontal flip	Enabled	Left and right breasts are mirror images
Vertical flip	Enabled	Extracted tiles carry no consistent anatomical orientation
Zoom range	$\pm 10\%$	Scale invariance across lesion sizes
Shear range	0.05	Mild geometric tolerance
Brightness range	[0.80, 1.20]	Detector exposure variability
Fill mode	Nearest	Avoids introducing border artefacts

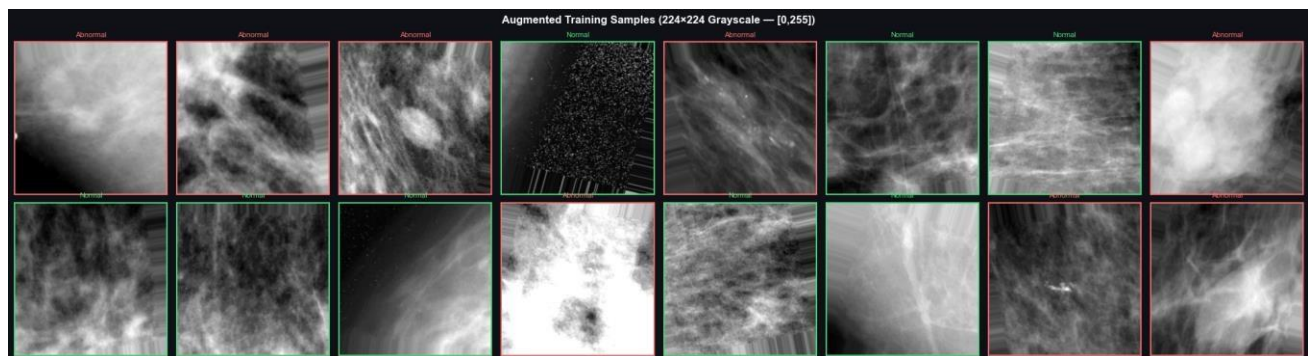


Fig. 4. Representative augmented training samples following the policy of Table III. Left group: Normal. Right group: Abnormal.

E. Data Splitting

Using stratified random splitting with a fixed seed of 42, the balanced set was split into training and validation subsets at an 80:20 ratio to achieve identical class balance in both partitions as indicated by Table II. Class label stratification was applied. The ramifications and limitations of this splitting strategy are explained in detail in Section V-F.

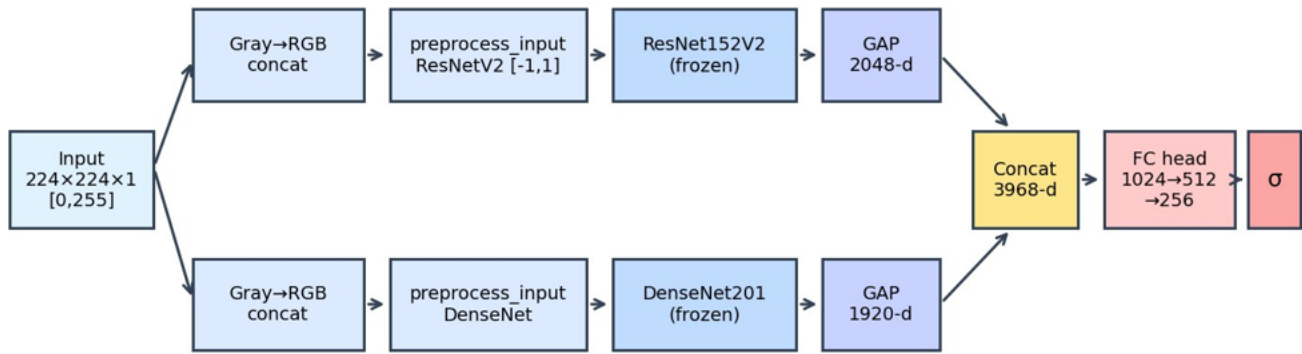
F. Model Architectures

Five single-backbone configurations were constructed identically. Each consists of the ImageNet-pretrained backbone with its classification top removed, followed by global average pooling and a shared dense classification head. The head comprises batch normalization, a 1024-unit dense layer with He uniform initialization, batch normalization, ReLU activation and dropout at rate 0.50; a 512-unit block with dropout 0.40; a 256-unit block with dropout 0.30; and a single sigmoid output unit.

A Dual-Branch Feature-Fusion Model of ResNet152V2 and DenseNet201 for Breast Abnormality Detection in Digitized Mammograms

The dual-branch fusion configuration, illustrated in Fig. 5, processes the same input through two parallel branches. Branch A applies grayscale-to-RGB replication, ResNet-V2 preprocessing and the frozen ResNet152V2 backbone, followed by global average pooling to a 2,048-dimensional vector. Branch B applies the same replication, DenseNet preprocessing and the frozen DenseNet201 backbone, pooling to a 1,920-dimensional vector. The two vectors are concatenated into a 3,968-dimensional joint representation and passed to the classification head described above.

It should be stated precisely that this is a feature-level fusion within a single network, not an ensemble in the conventional sense: there is no independent training of member models, no voting and no prediction averaging. The backbones themselves are unmodified reference implementations; the components original to this work are the fusion configuration, the in-graph dual preprocessing arrangement, and the shared descending-dropout head.



Total parameters: 81,397,313 | Trainable (Phase 1): 4,732,161

Fig. 5. Architecture of the dual-branch feature-fusion model. Both backbones remain frozen during Phase 1; the final 50 layers of each are unfrozen during Phase 2.

G. Mathematical Formulation

Let $x \in \mathbb{R}^{(224 \times 224 \times 1)}$ denote an input tile and $y \in \{0, 1\}$ its label, where 0 denotes Normal and 1 denotes Abnormal. Channel replication produces the three-channel tensor

$$x_3 = [x, x, x] \in \mathbb{R}^{(224 \times 224 \times 3)} \quad (1)$$

Each branch applies its backbone-specific preprocessing π and frozen convolutional function ϕ , producing a feature map

$$f = \phi(\pi(x_3)) \in \mathbb{R}^{(H \times W \times C)} \quad (2)$$

Global average pooling reduces this to a vector, where H and W denote the spatial dimensions of the final feature map and c indexes its channels:

$$g_c = (1/HW) \cdot \sum_h \sum_w f(h, w, c) \quad (3)$$

For the fusion model the two resulting vectors $g_R \in \mathbb{R}^{2048}$ and $g_D \in \mathbb{R}^{1920}$ are concatenated:

$$z = [g_R; g_D] \in \mathbb{R}^{3968} \quad (4)$$

The classification head applies three successive blocks, each consisting of a dense transformation with weights W_i and bias b_i , batch normalisation BN, ReLU activation and dropout with rate p_i :

$$h_i = \text{Dropout}_{\{p_i\}}(\text{ReLU}(\text{BN}(W_i h_{i-1} + b_i))) \quad (5)$$

with $h_0 = z$, $p_1 = 0.50$, $p_2 = 0.40$ and $p_3 = 0.30$. The predicted probability of the Abnormal class is

$$\hat{y} = \sigma(w_o^T h_3 + b_o), \quad \sigma(t) = 1 / (1 + e^{-t}) \quad (6) \text{ Training}$$

minimises the binary cross-entropy loss over N samples:

$$L = -(1/N) \sum_{n=1}^N [y_n \log \hat{y}_n + (1 - y_n) \log(1 - \hat{y}_n)] \quad (7)$$

Optimization uses Adam with the phase-specific learning rates given in Table IV. Discrete predictions are obtained by thresholding at $\tau = 0.5$, that is $\hat{y}_{\text{hard}} = 1$ if $\hat{y} \geq \tau$ and 0 otherwise.

A Dual-Branch Feature-Fusion Model of ResNet152V2 and DenseNet201 for Breast Abnormality Detection in Digitized Mammograms

H. Training Protocol

All configurations followed an identical two-phase schedule. In Phase 1 the backbone weights were frozen and only the classification head was trained, for 20 epochs with Adam at a learning rate of 10^{-3} . In Phase 2 the final 50 layers of each backbone were unfrozen and the entire network was trained for a further 30 epochs with Adam at a learning rate of 10^{-5} , a reduction of two orders of magnitude selected to avoid destroying the pretrained representations. Table IV records the complete configuration.

Table IV. Training Configuration, Identical Across All Reported Configurations

Parameter	Value
Framework	TensorFlow 2.10.1 / Keras
Environment	Python 3.10 (Anaconda), Windows, single GPU
Input shape	$224 \times 224 \times 1$
Batch size	16
Phase 1 (feature extraction)	20 epochs, Adam, learning rate 10^{-3} , backbone frozen
Phase 2 (fine-tuning)	30 epochs, Adam, learning rate 10^{-5} , final 50 backbone layers unfrozen
Loss function	Binary cross-entropy
Monitored metrics	Accuracy, AUC
Model checkpointing	Best validation AUC retained
Early stopping	Monitor validation loss, patience 10, best weights restored
Learning-rate schedule	ReduceLROnPlateau on validation loss, factor 0.5, patience 4, minimum 10^{-7}
Head initialization	He uniform
Regularization	Batch normalization; dropout 0.50 / 0.40 / 0.30
Decision threshold	0.50
Random seed	42 (partitioning and generator)

I. Evaluation Metrics

We report performance in terms of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) with pos=Abnormal class.

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (8)$$

$$Sensitivity = TP / (TP + FN) \quad (9)$$

$$Specificity = TN / (TN + FP) \quad (10)$$

$$Precision = TP / (TP + FP) \quad (11)$$

$$F1 = 2 \cdot (Precision \cdot Sensitivity) / (Precision + Sensitivity) \quad (12)$$

The AUC is computed over the entire range of decision thresholds. Both micro- and macro-averaged are reported; for ranking we take the macro-averaged value, which weights both classes equally (and is appropriate here as this is a balanced binary task). Because a false negative means that the abnormality is not detected, sensitivity is considered to be the clinically most important measure in this application.

Efficiency is further measured as a ratio of validation accuracy to the number of parameters (in million) — or accuracy per million parameter. This is not a standard measure and it is descriptive rather than inferential and is meant to clarify the bias– variance trade-off rather than rank them.

IV. RESULTS

All figures reported in this section were obtained on the 2,000-image validation partition described in Table II. As discussed in Section V-F, this partition also served for checkpoint selection and early stopping, and the reported values should therefore be interpreted as model-selection scores rather than as unbiased generalization estimates.

A Dual-Branch Feature-Fusion Model of ResNet152V2 and DenseNet201 for Breast Abnormality Detection in Digitized Mammograms

A. Comparative Performance

Table V reports the complete results for all six configurations, ordered by validation accuracy. Fig. 6 presents the same data graphically.

Table V. Comparative Performance Of The Six Configurations On The Validation Partition

Configuration	Val. Acc.	AUC	Sens.	Spec.	Prec.	F1	Params	Rank
Dual-branch fusion + (ResNet152V2 + DenseNet201)	0.9415	0.9845	0.9290	0.9580	0.9567	0.9427	81,397,313	1
ResNet152V2	0.9360	0.9835	0.9080	0.9700	0.9680	0.9370	61,101,569	2
VGG16	0.9315	0.9852	0.9510	0.9660	0.9655	0.9582	15,905,601	3
DenseNet201	0.9295	0.9806	0.9260	0.9340	0.9335	0.9297	20,960,321	4
Xception	0.9260	0.9801	0.8890	0.9550	0.9518	0.9193	23,631,401	5
MobileNetV2	0.8980	0.9814	0.8230	0.9760	0.9717	0.8912	4,238,401	6

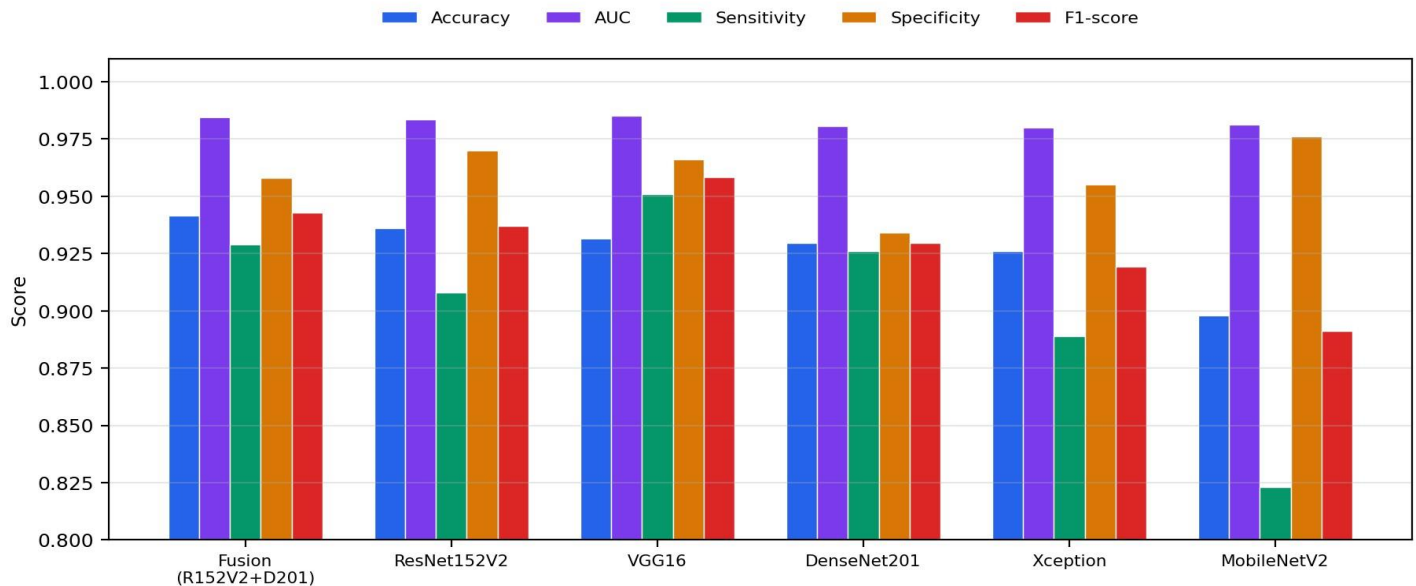


Fig. 6. Comparative performance across the five reported metrics for all six configurations.

The dual-branch fusion model achieved the highest validation accuracy at 0.9415, followed by ResNet152V2 at 0.9360 and VGG16 at 0.9315. The spread between the best and worst configuration was 4.35 percentage points of accuracy.

The ranking is not stable across metrics, and this is one of the more informative observations in the study. VGG16 ranked third on accuracy yet achieved the highest AUC (0.9852), the highest sensitivity (0.9510) and the highest F1-score (0.9582), while using the second-smallest parameter budget. ResNet152V2 achieved the second-highest specificity. MobileNetV2 ranked last on accuracy but achieved the highest specificity (0.9760) and the highest precision (0.9717), a direct consequence of its low sensitivity (0.8230): a model that predicts the Normal class more readily produces few false positives at the cost of missed abnormalities. Which configuration is preferred therefore depends on which metric is treated as primary.

B. Ablation of the Fusion Configuration

Since both constituent backbones were trained with the same protocol, we can directly measure rather than hypothesize about their fusion contribution. Table VI presents this ablation.

A Dual-Branch Feature-Fusion Model of ResNet152V2 and DenseNet201 for Breast Abnormality Detection in Digitized Mammograms

Table VI. Ablation of The Dual-Branch Fusion Configuration

Configuration	ResNet152V2	DenseNet201	Fusion head	Val. Acc.	AUC	F1	Params	Delta Acc.
Branch A alone	Y	-	-	0.9360	0.9835	0.9370	61,101,569	reference
Branch B alone	-	Y	-	0.9295	0.9806	0.9297	20,960,321	-0.0065
Dual-branch fusion	Y	Y	Y	0.9415	0.9845	0.9427	81,397,313	+0.0055

The fusion configuration exceeded its stronger constituent branch by 0.55 percentage points of accuracy (0.9415 against 0.9360) and its weaker constituent by 1.20 points. It also achieved the higher AUC (0.9845 against 0.9835) and the higher F1-score (0.9427 against 0.9370). The gain is therefore consistent in direction across three independent metrics, which is more informative than the magnitude of any single one.

The gain must nonetheless be weighed against its cost. Fusion added 20.30 million parameters relative to Branch A alone, an increase of 33.2%, in exchange for 0.55 percentage points of accuracy — equivalent to eleven additional correct classifications out of two thousand. Because only a single training run was performed per configuration, no confidence interval is available and the possibility that a difference of this magnitude reflects run-to-run variation cannot be excluded. This is addressed in Section V-F.

C. Training Behaviour and Error Structure of the Proposed Model

Fig. 7 presents the accuracy, loss and AUC trajectories of the dual-branch fusion model across both training phases. The general two-stage profile is evident: fast improvement during frozen-feature extraction and stable performance thereafter, followed by another jump when a partial fine-tuning starts. The learning-rate scheduler reduced the rate at least once, and early stopping restored the best weights.



Fig. 7. Training and validation accuracy, loss and AUC for the dual-branch fusion model.

Fig. 8 presents the confusion matrix of the fusion model in absolute counts and row-normalized percentages, and Fig. 9 the corresponding per-class precision, recall and F1-score. Fig. 10 presents the receiver operating characteristic curves with per-class, micro- and macro-averaged curves and a magnified view of the upper-left region.

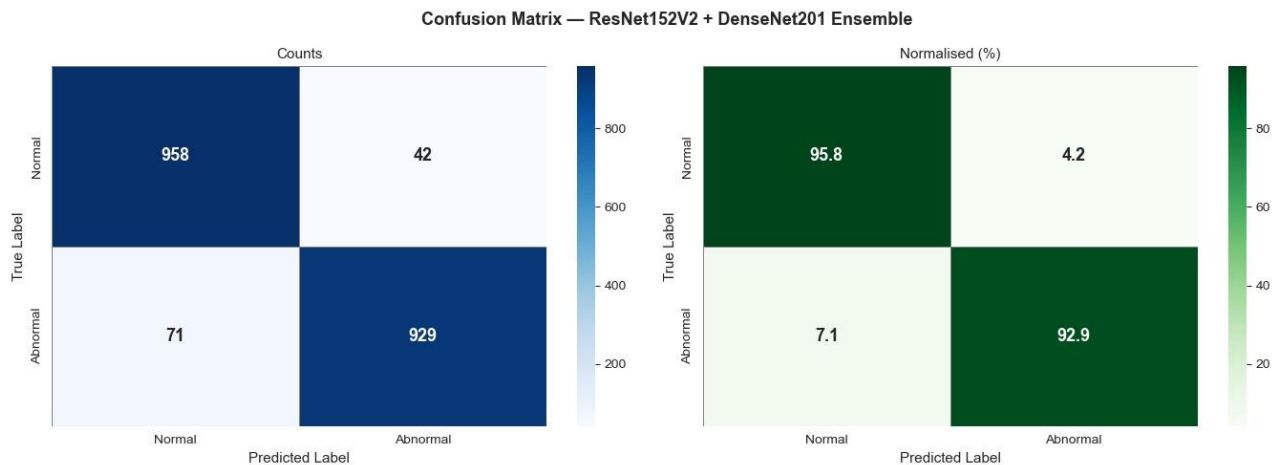


Fig. 8. Confusion matrix for the dual-branch fusion model in absolute counts (left) and row-normalised percentages (right).

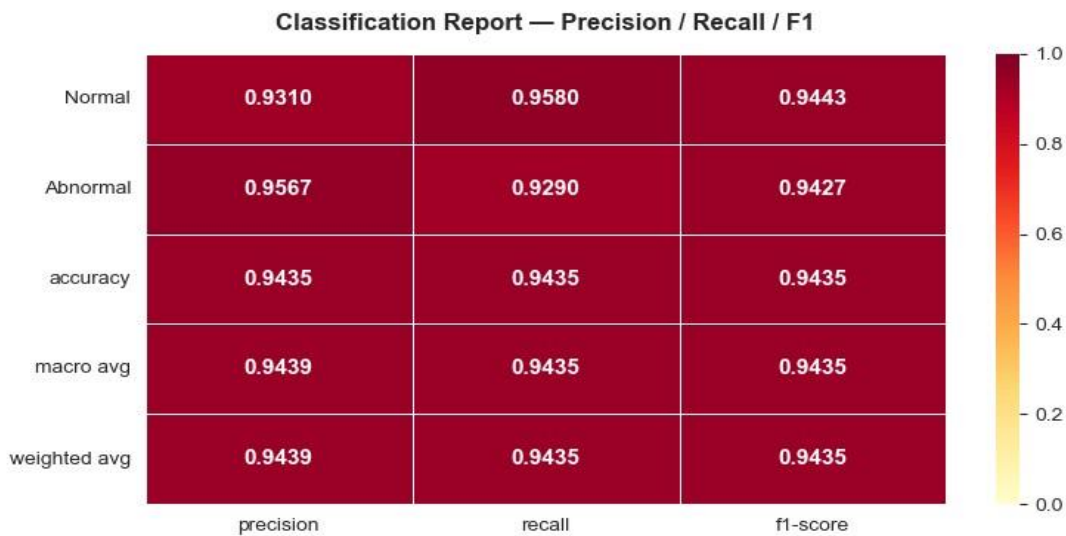


Fig. 9. Per-class precision, recall and F1-score for the dual-branch fusion model.

The per-class view in Fig. 9 shows that the two classes are recognized with comparable but not identical quality. For the Abnormal class the model attains a precision of 0.9567 against a recall of 0.9290, whereas for the Normal class the relationship inverts, with a precision of 0.9310 against a recall of 0.9580. The resulting F1-scores are close, 0.9427 for Abnormal and 0.9443 for Normal. This behaviour is a direct reflection of the threshold behaviour described in Section IV-D: since it labels Normal more liberally, it also has higher recall against all normal tiles and thus gets captured into a greater share but requires strict accuracy on its abnormal predictions (thus providing better precision for the Abnormal from this perspective). The relatively similar values of the two F1-scores show that neither class suffers a consistent disadvantage overall, and that the imbalance represents one of error type only, not ability overall.

A Dual-Branch Feature-Fusion Model of ResNet152V2 and DenseNet201 for Breast Abnormality Detection in Digitized Mammograms

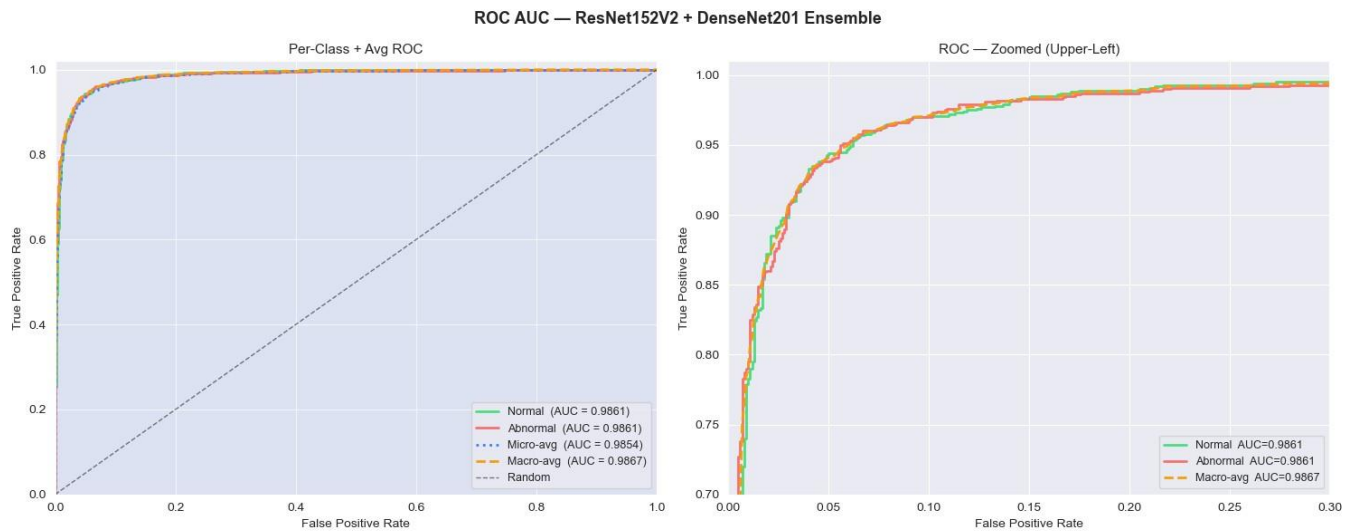


Fig. 10. ROC curves for the dual-branch fusion model.

The ROC analysis in Fig. 10 quantifies the brake on predictive accuracy, irrespective of a given decision threshold. The per-class curves are virtually coincident, as expected behaviour in the case of a balanced binary problem, since each one is essentially mirroring part of the same underlying decision function, and both approach the upper-left corner sharply. The magnified panel shows the relevant operating region: a true-positive rate greater than 0.90 is achieved at a false-positive rate of around 0.05, and the curve remains near unity thereafter. This observation justifies the strong quality of ranking throughout the whole range of thresholds reported in Table V as a macro-averaged AUC (0.9845). The pragmatic interpretation is that the sensitivity shortfall captured in Section \ref{sec:sensitivity-shortfall} should not be viewed as a limitation on the learned representation, but rather on where the operating point was pinned: the curvature of this plot suggests that there is significant potential to gain sensitivity if we were willing to go below a threshold of 0.5, at a much smaller proportional cost in specificity.

D. Sensitivity and Specificity Balance

Table V shows a consistent pattern in that specificity, with the exception of DenseNet201, is always greater than sensitivity (often by a significant margin). The gap reaches 15.30 percentage points for MobileNetV2 (0.8230 against 0.9760) and 6.20 points for ResNet152V2 (0.9080 against 0.9700). Only VGG16 approaches balance, with 0.9510 sensitivity against 0.9660 specificity.

At the fixed decision threshold of 0.5, the models are therefore biased toward predicting the Normal class, and the missed abnormality rate exceeds the false-alarm rate. This has direct clinical relevance, since in a screening context a false negative carries greater consequence than a false positive, which leads to recall and further imaging rather than to a missed cancer. Because all six configurations achieved AUC above 0.98, ranking quality is high and the threshold can be lowered to trade specificity for sensitivity at modest cost. The implication is examined in Section V-D.

E. Accuracy-Complexity Relationship

Fig. 11 plots validation accuracy against total parameter count. The relationship is weak and clearly non-monotonic. The two largest configurations occupy first and second place, but VGG16 at 15.91 million parameters falls only 1.00 percentage point behind the 81.40-million-parameter fusion model while using less than a fifth of the parameters, and leads it on AUC, sensitivity and F1-score.

A Dual-Branch Feature-Fusion Model of ResNet152V2 and DenseNet201 for Breast Abnormality Detection in Digitized Mammograms

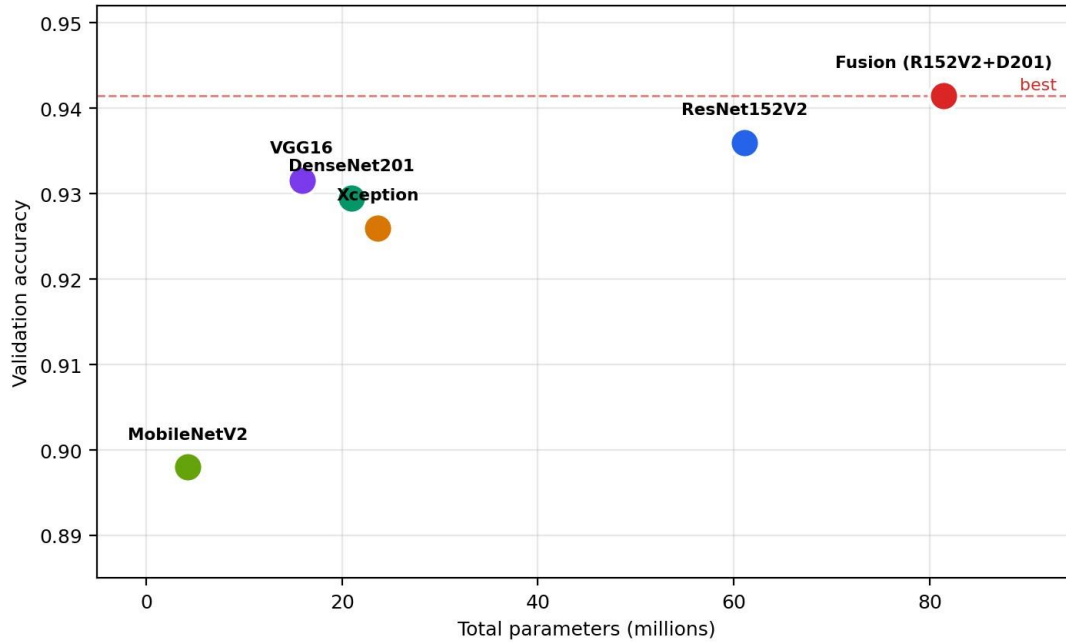


Fig. 11. Validation accuracy against total parameter count. The dashed line marks the best observed accuracy.

Fig. 12 presents accuracy per million parameters. MobileNetV2 is a clear outlier, achieving 0.2119 accuracy per million parameters against 0.0586 for VGG16 — a factor of 3.6 — and 0.0116 for the fusion model, a factor of 18.3. Table VII summarizes the efficiency comparison.

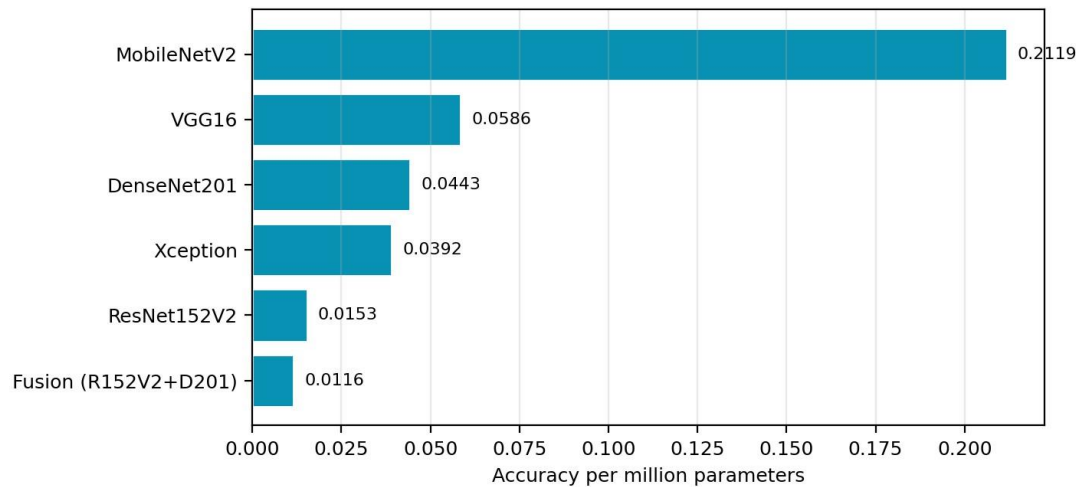


Fig. 12. Accuracy per million parameters for the six configurations.

Table VII. Computational Characteristics

Configuration	Params (M)	Val. Acc.	Acc. per M params	Relative to MobileNetV2
MobileNetV2	4.24	0.8980	0.2119	1.00x
VGG16	15.91	0.9315	0.0586	0.28x
DenseNet201	20.96	0.9295	0.0443	0.21x
Xception	23.63	0.9260	0.0392	0.18x
ResNet152V2	61.10	0.9360	0.0153	0.07x
Dual-branch fusion	81.40	0.9415	0.0116	0.05x

A Dual-Branch Feature-Fusion Model of ResNet152V2 and DenseNet201 for Breast Abnormality Detection in Digitized Mammograms

One additional measurement is available for the fusion configuration: prediction over the 2,000 validation images required approximately 30 seconds at batch size 16, corresponding to roughly 15 ms per image, and the serialized model occupied 488.0 MB on disk. Equivalent measurements were not recorded for the remaining configurations and are identified as an item for completion in Section VI.

V. DISCUSSION

A. Principal Findings

Three findings emerge. First, dual-branch feature fusion produced a measurable improvement over both of its constituent backbones, exceeding the stronger of the two by 0.55 percentage points of accuracy and doing so consistently across accuracy, AUC and F1-score. Second, the improvement is small relative to its cost: 20.30 million additional parameters purchased eleven additional correct classifications out of two thousand. Third, and most consequential for how such comparisons should be read, the preferred configuration changes with the metric: fusion leads on accuracy, VGG16 on AUC, sensitivity and F1-score, and MobileNetV2 on specificity, precision and parameter efficiency. No configuration dominates across all metrics.

B. Comparison with Previous Studies

These findings should be compared with prior work cautiously, because the task formulation differs. The five studies reviewed in Section II all perform benign-versus-malignant classification on curated regions of interest, whereas the present study performs normal-versus-abnormal classification on extracted tiles. The latter is generally the easier discrimination, and absolute accuracies are consequently not directly comparable.

With that qualification, the results are consistent with the literature in magnitude and direction. Meva et al. [16] reported $93.7\% \pm 1.0\%$ for an attention-augmented three-model ensemble on CBIS-DDSM; the present fusion result of 94.15% sits marginally above that interval on a different and easier task, which is the expected relationship rather than evidence of superiority. Das et al. [14] reported 89.2% for fine-tuned deep networks on CBIS-DDSM and Puttegowda et al. [15] reported 87.08%, both below the range observed here, again consistent with the task difference.

The agreement on fusion is more informative than the agreement on absolute values. Meva et al. [16] found that combining architectures improved on their individual baselines, and the present ablation reaches the same conclusion under a protocol in which both constituent backbones were trained identically to the fusion model. Against this, Puttegowda et al. [15] concluded that preprocessing and class balancing influenced outcomes more strongly than backbone choice, and Priyanka et al. [11] found no statistically significant differences among nine architectures under patient-aware validation with corrected testing. The present observation that six configurations of widely differing sophistication span only 4.35 percentage points, and that the ordering changes with the metric, is more compatible with the latter position than with the view that architecture selection is decisive.

C. Mechanistic Interpretation

Two observations invite explanation: why fusion helped, and why it helped so little.

The gain is consistent with complementary error patterns between the two branches. ResNet152V2 and DenseNet201 differ in how features propagate — residual addition against dense concatenation — and their pooled representations are unlikely to be identical. When they are concatenated, the classification head has access to both, and examples mis-classified by one branch can be recoverable if the other encodes discriminative evidence. This account is supported by the direction of effect observed in both cases: sensitivity improved compared to ResNet152V2 (0.9290 against 0.9080) but specificity given up (0.9580 against 0.9700), which coincides with what would be expected if evidence for the Abnormal class was contributed by that branch of DenseNet201.

The same is true of the small size. Since both of these backbones are deep hierarchies pretrained on the same ImageNet distribution, it is very likely that their pooled representations encode substantially overlapping information; concatenation thus increases the head input space from 2,048 to 3,968 dimensions without a proportionate increase in discriminative signal. The main reason for the small gains is that under this account most of the newly available capacity appears to be spurious redundancy at massive overhead. This means that the hypothesis is directly testable as a correlation or centred kernel alignment between the two pooled vectors — we identify this analysis for future work in Section VI.

A third observation deserves comment. The simplest architecture among these, i.e. VGG16, performed the highest AUC, sensitivity and F1-score based on 15.91 million of parameters. Although backbone weights were frozen in Phase 1, with only final layers adapted

A Dual-Branch Feature-Fusion Model of ResNet152V2 and DenseNet201 for Breast Abnormality Detection in Digitized Mammograms

in Phase 2, the total count of trainable parameters was restricted to 1.19 (DNet-121)–4.73 million (CommonNetB0) against a dataset size of 8k training images. A simple but wide feature hierarchy under such limited adaptation may transfer more readily than a deep one, and similar evidence appears in the broader transfer-learning literature [3].

D. Clinical and Practical Implications

The sensitivity-specificity imbalance documented in Section IV-D has direct practical significance. In a screening application the cost of a false negative — a missed abnormality — is not symmetric with that of a false positive, which results in recall and supplementary imaging. Five of the six configurations exhibit specificity exceeding sensitivity at the default threshold of 0.5, in the case of MobileNetV2 by 15.30 percentage points. Because all configurations achieved AUC above 0.98, the ranking quality is high and the operating point can be moved: a deployment configuration would select the threshold from a target sensitivity rather than defaulting to 0.5. On this criterion VGG16, which is closest to balance and attains the highest AUC, sensitivity and F1-score, may be preferable to the accuracy-ranked leader.

The efficiency results carry a second implication. MobileNetV2 reached 0.8980 accuracy with 4.24 million parameters, 4.35 percentage points below the fusion model while using 19.2 times fewer parameters. Its very high precision and specificity make it a plausible triage component for filtering clearly normal examinations, though its low sensitivity rules it out as a standalone detector. More generally, the results show that accuracy reported without parameter cost is insufficient information for a deployment decision, and that reporting the two jointly costs nothing.

E. Strengths

The study has four methodological strengths. All configurations share one preprocessing pipeline, eliminating preprocessing as a confounder. The augmentation policy, classification head, loss function and callback configuration are common across configurations. Both constituent backbones of the fusion model are evaluated independently, providing a complete rather than a

partial ablation — a control that much of the fusion literature omits. Finally, the pipeline is documented at the level of individual hyperparameters and released in full, permitting exact reproduction.

F. Limitations

Several limitations must be stated plainly, as they bound the conclusions that may be drawn.

No independent test partition. The 2,000-image validation set was used for early stopping, checkpoint selection on validation AUC, and learning-rate scheduling, in addition to serving as the reported evaluation set. All reported figures are therefore optimistically biased model-selection scores rather than unbiased generalization estimates. The comparison between configurations remains internally valid, since all were treated identically, but the absolute values should not be interpreted as expected performance on unseen data.

Image-level rather than source-level partitioning. The partition was performed on a flat image array with stratification on class label only. The DDSM TFRecord distribution comprises tiles extracted from mammographic examinations, and no examination or patient identifier was available in the parsed feature set. Tiles originating from the same source examination may therefore appear in both partitions. Published quantifications of this effect report accuracy inflation of five to thirty percent [8] and up to fifty-five percent in volumetric settings [9]. This is the most serious limitation of the study, and absolute performance figures should be regarded as upper bounds until the analysis is repeated with grouped partitioning.

Single training run per configuration. Each configuration was trained once with a fixed seed. No variance, confidence interval or significance test is available. This bears directly on the central claim: the fusion gain of 0.55 percentage points over ResNet152V2 corresponds to eleven images out of two thousand, and cannot be distinguished from stochastic variation on the basis of a single run, and the claim that fusion improves on its constituent backbones should be regarded as supported in direction but not established in magnitude.

Protocol equalization should be verified for all configurations. The comparison rests on the premise that architecture was the only varying factor. Any configuration trained with a different batch size, fine-tuning learning rate or number of unfrozen layers is not protocol-matched to the others, and its position in Table V would reflect tuning as well as architecture. This should be confirmed against the training logs before the ranking is relied upon.

No external validation. All results derive from a single dataset. No evaluation on INbreast, CBIS-DDSM or a full-field digital mammography cohort was performed, so the transferability of these findings is untested.

A Dual-Branch Feature-Fusion Model of ResNet152V2 and DenseNet201 for Breast Abnormality Detection in Digitized Mammograms

The mammograms in the DDSM dataset are obtained by scanning old film-based images into digital files; they are not captured by the full-field digital mammography and tomosynthesis devices that are currently the mainstream for clinical breast screening. The noise characteristics and dynamic range of the images produced by these two types of new devices are different from those scanned from old films. Therefore, the research conclusions drawn from this old dataset cannot necessarily be directly applied to the new image acquisition hardware currently in use.

This dataset adopts a category-quotaed convenience sampling method. When constructing the dataset, the creators extracted images from a larger source data pool, and set a maximum limit of 5,000 images for each category in advance. At that time, there were five packaged partitioned datasets available, and only four of them were used to meet the quota of 5,000 images per category, leaving the fifth partition unused. In the end, the number of samples for all categories was exactly equal, achieving precise category balance, but there was a problem: these samples were not randomly selected from all source images.

The task framework set for this study is relatively simplistic. It only performs the simple classification of distinguishing between normal and abnormal cases, and the difficulty of this task in clinical settings is far lower than that of two more demanding tasks: one is distinguishing whether a lesion is benign or malignant, and the other is classifying images according to the BI-RADS standard. Even if excellent results are achieved on this basic task, it cannot be used as evidence that this method is already capable of handling those more difficult tasks in real clinical scenarios.

Absence of demographic characterization. By using this distribution, no demographic metadata were available and the HR performance in sub-groups partitioned across breast density, age or ethnicity could not be analyzed. A consistent criticism of this literature, identified by systematic reviews, is the lack of attention to subgroup reporting with poorly representative demographics [2].

G. Generalizability

Since all configurations were treated similarly in this protocol, internal comparisons are probably comparatively robust (though the limitations outlined above apply uniformly to all configurations). While the specific values are easily out-dated, the qualitative conclusions — that fusion gives a modest gain over its components and that architecture choice is metric-indifferent — are more robust. The absolute performance figures are considerably less generalizable, being bounded by the validation-set evaluation, the partitioning strategy and the single dataset. Transfer to full-field digital mammography, to other populations, and to benign-versus-malignant discrimination is untested and should not be assumed.

VI. CONCLUSION AND FUTURE WORK

This study compared six configurations for normal-versus-abnormal classification of digitized mammographic tiles from DDSM under a protocol in which preprocessing, augmentation, classification head, loss function and callbacks were held constant. On a class-balanced set of 10,000 tiles, a dual-branch model fusing the pooled feature vectors of ResNet152V2 and DenseNet201 achieved the highest validation accuracy at 0.9415 with an AUC of 0.9845, exceeding its stronger constituent backbone by 0.55 percentage points and its weaker constituent by 1.20 points, with the gain consistent in direction across accuracy, AUC and F1score. VGG16 achieved the highest AUC, sensitivity and F1-score using 15.91 million parameters. MobileNetV2 reached 0.8980 accuracy with 4.24 million parameters, the highest accuracy-per-parameter ratio by a factor of 3.6.

Two conclusions follow. Dual-branch feature fusion produces a real but modest gain over its constituent backbones, at a parameter cost that must be weighed explicitly rather than assumed to be worthwhile. And because no configuration dominated across all metrics, the choice of architecture cannot be separated from the choice of which metric matters for the intended application. The principal limitation is that evaluation was performed on the same partition used for model selection, with imagelevel rather than source-level partitioning, so absolute figures should be treated as upper bounds.

Future work follows directly from these limitations. The highest priority is to repartition the data with examination- or patientlevel grouping and to reserve an independent test set, then to repeat the comparison and quantify the difference between the two partitioning strategies on the same data — a measurement that is itself an identified gap in the mammographic literature. Second, each configuration should be repeated across multiple random seeds so that confidence intervals and corrected pairwise significance tests can be reported; this is necessary before the fusion gain can be described as established rather than indicated. Third, the feature-redundancy hypothesis advanced in Section V-C should be tested directly by measuring representational similarity between the two pooled backbone vectors, which would explain the small magnitude of the fusion gain. Fourth, threshold optimization and calibration

A Dual-Branch Feature-Fusion Model of ResNet152V2 and DenseNet201 for Breast Abnormality Detection in Digitized Mammograms

analysis would address the sensitivity-specificity imbalance documented in Section IV-D. Fifth, external validation on INbreast and on a full-field digital cohort would establish how far these findings transfer. Finally, gradient-based explanation methods reviewed against annotated lesions would address the interpretability expectations of clinical deployment.

APPENDIX A

Per-Configuration Training Behaviour

Figs. 13 through 17 present the accuracy, loss and AUC trajectories for the five single-backbone configurations, using the same three-panel format as Fig. 7. All exhibit the characteristic two-stage profile, with the vertical marker denoting the transition from frozen feature extraction to partial fine-tuning of the final backbone layers.



Fig. 13. Training history for ResNet152V2.

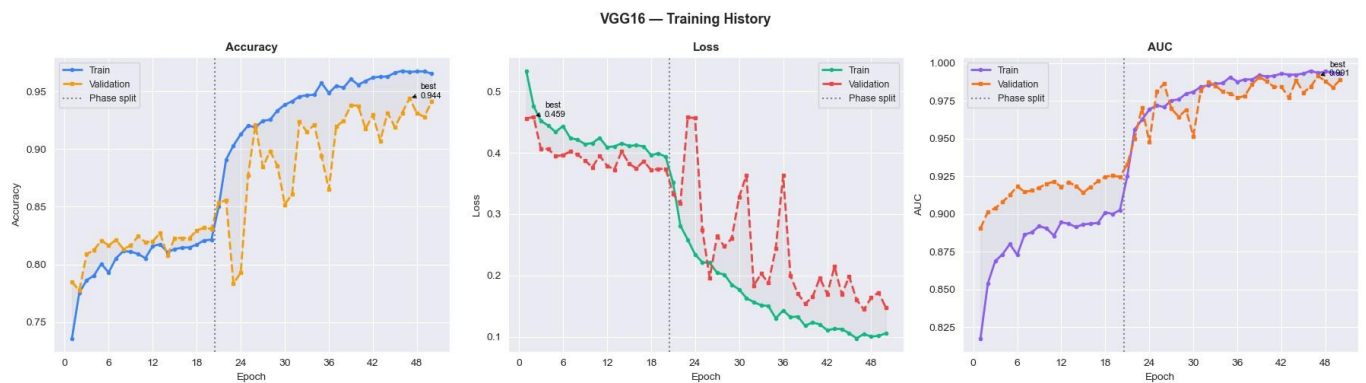


Fig. 14. Training history for VGG16.

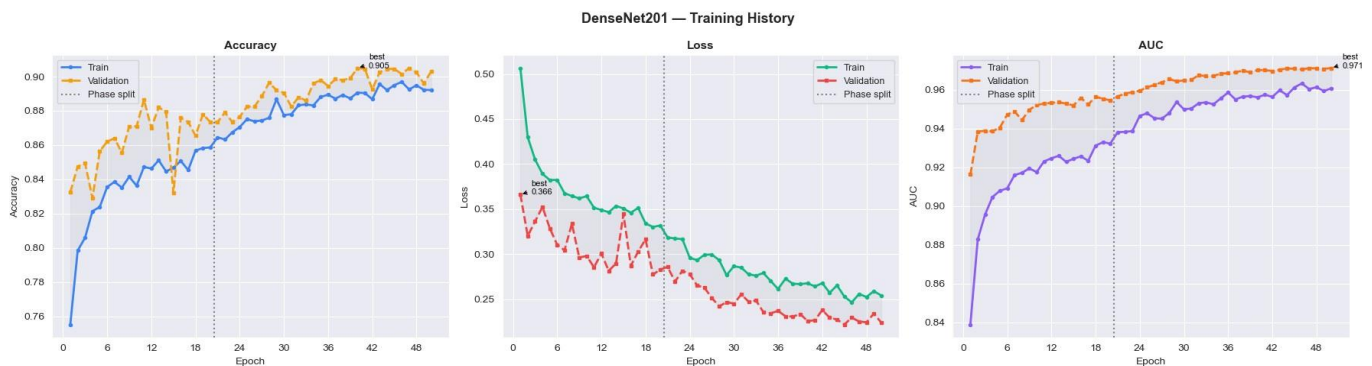


Fig. 15. Training history for DenseNet201.

A Dual-Branch Feature-Fusion Model of ResNet152V2 and DenseNet201 for Breast Abnormality Detection in Digitized Mammograms



Fig. 16. Training history for Xception.



Fig. 17. Training history for MobileNetV2.

APPENDIX B

Per-Configuration Error Structure

Figs. 18 through 22 present the confusion matrices for the five single-backbone configurations in absolute counts and rownormalized percentages, in the same format as Fig. 8.

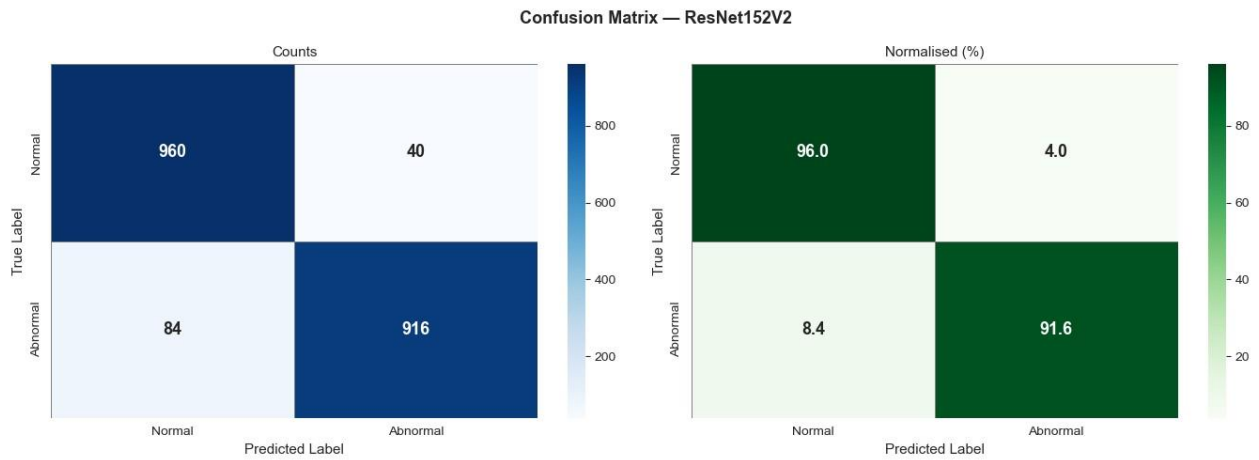


Fig. 18. Confusion matrix for ResNet152V2.

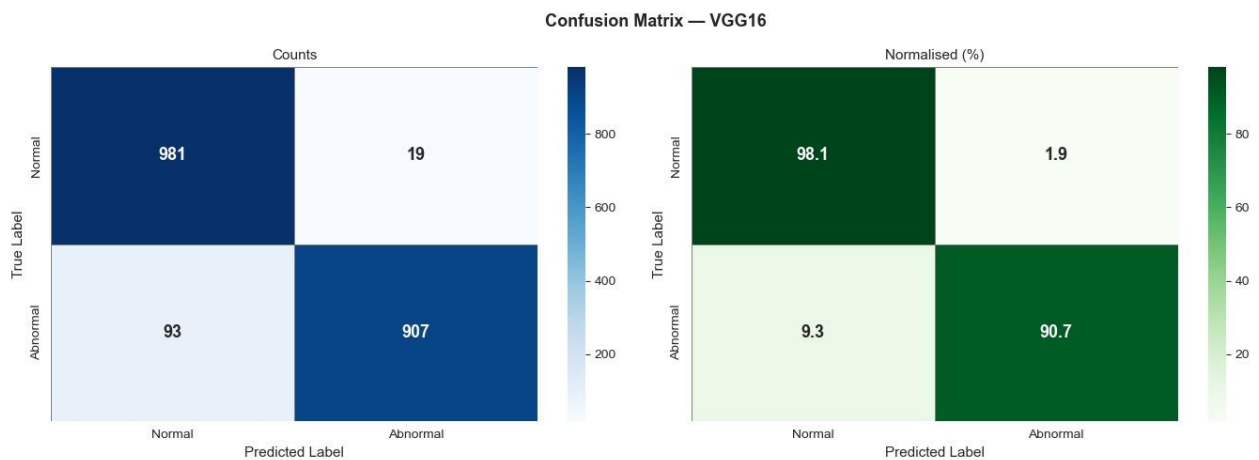


Fig. 19. Confusion matrix for VGG16.

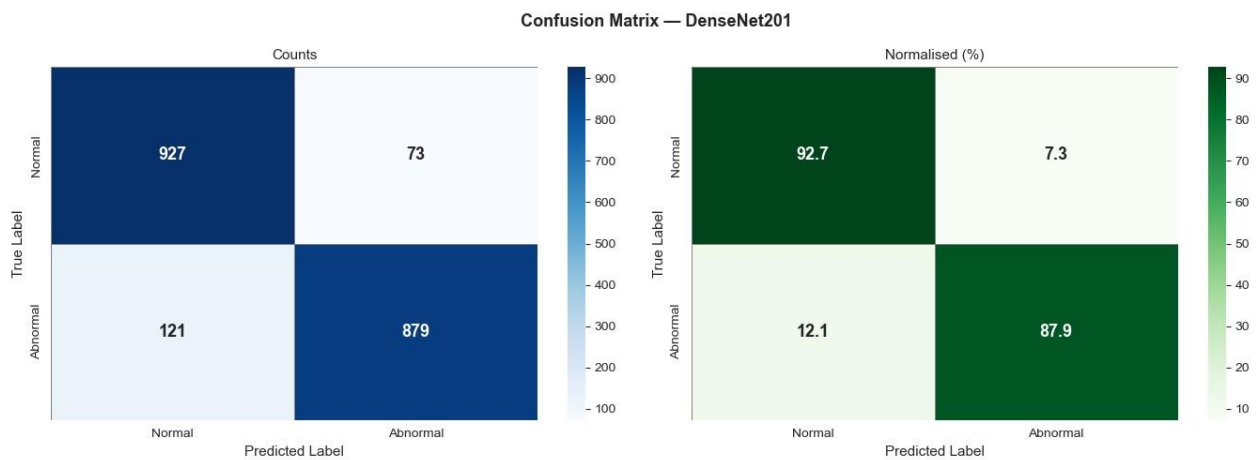


Fig. 20. Confusion matrix for DenseNet201.

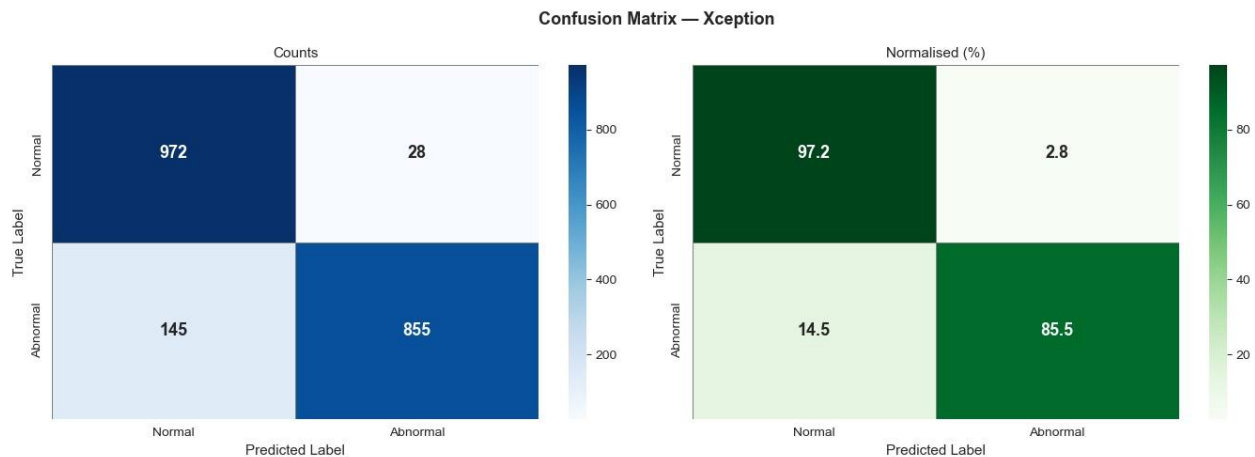


Fig. 21. Confusion matrix for Xception.

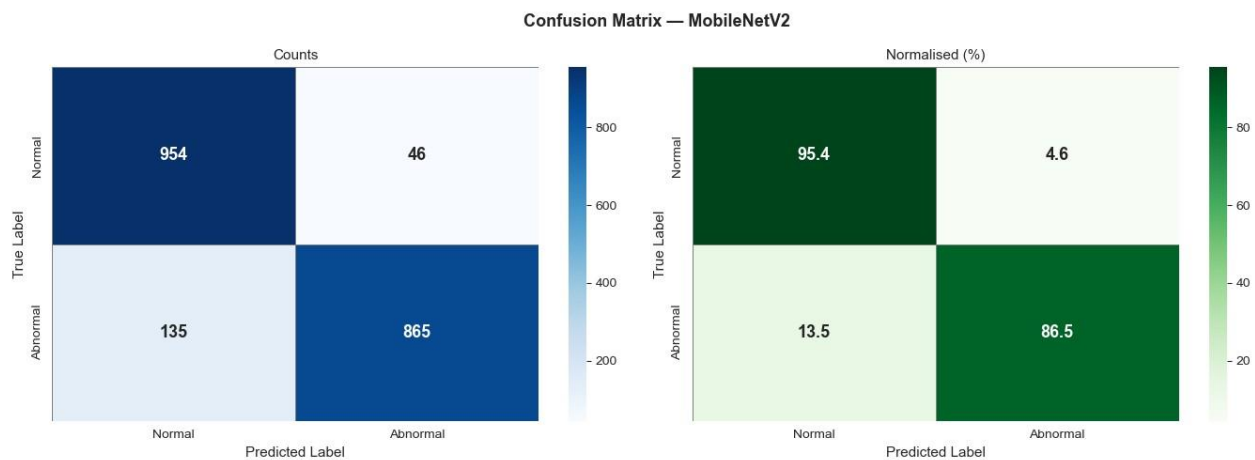


Fig. 22. Confusion matrix for MobileNetV2.

REFERENCES

- 1) L. Shen, L. R. Margolies, J. H. Rothstein, E. Fluder, R. McBride, and W. Sieh, "Deep learning to improve breast cancer detection on screening mammography," *Sci. Rep.*, vol. 7, no. 1, Art. no. 4165, 2017.
- 2) S. Gao, J. Liu, L. Li, D. Yang, Y. Miao, X. Zhang, Q. Han, Y. Shi, J. Wu, and K. Zhang, "Application of deep learning technology in breast cancer: a systematic review of segmentation, detection, and classification approaches," *BioMed. Eng. OnLine*, vol. 25, 2026, doi: 10.1186/s12938-025-01502-5.
- 3) H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, "Transfer learning for medical image classification: a literature review," *BMC Med. Imaging*, vol. 22, Art. no. 69, 2022.
- 4) M. A. Morid, A. Borjali, and G. Del Fioli, "A scoping review of transfer learning research on medical image analysis using ImageNet," *Comput. Biol. Med.*, vol. 128, Art. no. 104115, 2021.
- 5) S. Mohapatra, S. Muduly, S. Mohanty, J. V. R. Ravindra, and S. N. Mohanty, "Evaluation of deep learning models for detecting breast cancer using mammograms," *Sustain. Oper. Comput.*, vol. 3, pp. 296–302, 2022.
- 6) P. Oza, P. Sharma, and S. Patel, "Breast lesion classification from mammograms using deep neural network and test-time augmentation," *Neural Comput. Appl.*, 2023.
- 7) A. Jaamour, C. Myles, A. Patel, S.-J. Chen, L. McMillan, and D. Harris-Birtill, "A divide and conquer approach to maximise deep learning mammography classification accuracies," *PLOS ONE*, vol. 18, no. 5, e0280841, 2023.
- 8) I. E. Tampu, A. Eklund, and N. Haj-Hosseini, "Inflation of test accuracy due to data leakage in deep learning-based classification of OCT images," *Sci. Data*, vol. 9, Art. no. 580, 2022.

A Dual-Branch Feature-Fusion Model of ResNet152V2 and DenseNet201 for Breast Abnormality Detection in Digitized Mammograms

- 9) E. Yagis, S. W. Atnafu, A. García Seco de Herrera, et al., "Effect of data leakage in brain MRI classification using 2D convolutional neural networks," *Sci. Rep.*, vol. 11, Art. no. 22544, 2021.
- 10) J. Wen, E. Thibeau-Sutre, M. Diaz-Melo, et al., "Convolutional neural networks for classification of Alzheimer's disease: overview and reproducible evaluation," *Med. Image Anal.*, vol. 63, Art. no. 101694, 2020.
- 11) V. Priyanka, M. Narendra, and T. D. Rao, "A patient-aware benchmarking of CNN and transformer architectures for breast cancer histopathology classification," *Front. Digit. Health*, vol. 8, 2026, doi: 10.3389/fdgth.2026.1752938.
- 12) D. A. Ragab, O. Attallah, M. Sharkas, J. Ren, and S. Marshall, "A framework for breast cancer classification using multi-DCNNs," *Comput. Biol. Med.*, vol. 131, Art. no. 104245, 2021.
- 13) S. Chakravarthy, N. Bharanidharan, and H. Rajaguru, "Multi-class breast cancer classification using CNN features hybridization," *Int. J. Comput. Intell. Syst.*, vol. 17, 2024.
- 14) H. S. Das, A. Das, A. Neog, S. Mallik, K. Bora, and Z. Zhao, "Breast cancer detection: shallow convolutional neural network against deep convolutional neural networks based approach," *Front. Genet.*, vol. 13, 2023.
- 15) K. Puttegowda, V. V. R. K. H. S, P. H. L, and S. K. V, "Accurate Mammogram classification for Breast Cancer detection using Transfer Learning-based CNN models," *Journal of Integrated Science and Technology*, vol. 14, no. 4, pp. 1559–1559, Mar. 2026, doi: 10.62110/sciencein.jist.2026.v14.1559.
- 16) D. T. Meva, K. Popat, H. Kukadiya, and N. Jariwala, "Deep learning-based breast cancer detection with customized ensemble attention," *Sci Rep*, vol. 16, no. 1, p. 20022, Apr. 2026, doi: 10.1038/s41598-026-50563-6.
- 17) D. A. Ragab, M. Sharkas, S. Marshall, and J. Ren, "Breast cancer detection using deep convolutional neural networks and support vector machines," *PeerJ*, vol. 7, e6201, 2019.
- 18) L. Bouzar-Benlabiod, K. Harrar, L. Yamoun, M. Y. Khodja, and M. A. Akhloufi, "A novel breast cancer detection architecture based on a CNN-CBR system for mammogram classification," *Comput. Biol. Med.*, vol. 163, Art. no. 107133, 2023.
- 19) K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015.
- 20) K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 630–645.
- 21) G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 4700–4708.
- 22) F. Chollet, "Xception: deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 1251–1258.
- 23) M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 4510–4520.
- 24) C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 2818–2826.
- 25) J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: a large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2009, pp. 248–255.
- 26) D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015.
- 27) S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 448–456.
- 28) N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, 2014.
- 29) K. Freeman, J. Geppert, C. Stinton, et al., "Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy," *BMJ*, vol. 374, n1872, 2021.
- 30) D. Abdelhafiz, C. Yang, R. Ammar, and S. Nabavi, "Deep convolutional neural networks for mammography: advances, challenges and applications," *BMC Bioinformatics*, vol. 20, Art. no. 281, 2019.