

Comparative Deep Learning and Explainable AI Framework for Medical Image Disease Classification: Baseline CNN versus ResNet50, MobileNetV2, and EfficientNet-B0

¹Kaynaat Hussain, ²Priya Sen, ³Kunal Batra

¹Hussainkaynaat@gmail.com, ²Priya.sen@sageuniversity.in, ³Kunal.batra@sageuniversity.in

^{1, 2, 3} Computer Science & Engineering, Institute of Engineering & Technology, Sage University Indore

Abstract—Reliable automated classification of medical images requires models that are accurate, efficient, and open to scrutiny. This paper reports a comparative evaluation of a baseline convolutional neural network (CNN) and three transfer-learning architectures, ResNet50, MobileNetV2, and EfficientNet-B0, on a two-class medical image classification task comprising 4,708 training, 524 validation, and 624 test images. The training and validation partitions are imbalanced (about 74% Class 1), and the test partition has a different class mix (62.5% Class 1). On the test set, the baseline CNN reached 90.87% accuracy, ResNet50 93.43%, EfficientNet-B0 94.23%, and MobileNetV2 95.03%, with weighted F1-scores of 90.70%, 93.31%, 94.13%, and 94.97%, respectively. Confusion-matrix analysis shows that most residual errors in every model are Class 0 images predicted as Class 1, and class-wise recall for Class 0 (80.77%–88.03%) is consistently lower than for Class 1 (96.92%–99.49%). MobileNetV2 achieved the highest reported accuracy with 2,226,434 parameters, about one-tenth of ResNet50's 23,512,130. Within the evaluated dataset, these observations indicate that compact pre-trained architectures can be competitive, while aggregate accuracy conceals asymmetric class behaviour. Dataset identity, preprocessing, augmentation, and most training details were not specified in the supplied material, and no explainability outputs were generated; a planned post-hoc XAI evaluation is therefore outlined as a future extension rather than reported as a result.

Index Terms—medical image classification, deep learning, transfer learning, ResNet50, MobileNetV2, EfficientNet-B0, class imbalance, confusion matrix, explainable AI, Grad-CAM

I. INTRODUCTION

Deep learning has become a principal approach to medical image classification, with applications reported for brain MRI, fundus photography, histopathology, and chest radiography [1]–[5]. In practice, researchers must choose among model families that differ markedly in size and design, and the choice is rarely guided by controlled comparison on a common dataset. Heavy residual networks are widely used as backbones [6], whereas lightweight networks are attractive when computation or memory is limited. Moreover, reporting accuracy alone provides an incomplete picture of diagnostic behaviour, particularly when classes are imbalanced.

This paper presents a comparative implementation study motivated by the companion review on post-hoc explainable AI (XAI) for medical imaging. The review concluded that accuracy benchmarking and explanation analysis are frequently separated and that explanation research seldom accounts for class imbalance or architectural differences. The present study addresses the benchmarking half of that agenda: it evaluates four models under a common test partition using accuracy, weighted and class-wise precision, recall, F1-score, confusion matrices, and parameter counts, and it then specifies how post-hoc explanation analysis should be added.

Contributions. (i) A side-by-side evaluation of a baseline CNN and three pre-trained architectures on one test set; (ii) class-wise and confusion-matrix analysis that exposes

asymmetric errors hidden by aggregate scores; (iii) an accuracy–size comparison across the four models; and (iv) a transparent statement of what the supplied material does and does not contain, together with a planned XAI evaluation. No claim of clinical readiness, statistical superiority, or experimental XAI performance is made.

II. RELATED WORK

Transfer learning is a recurrent strategy where labelled medical data are limited. Fine-tuned EfficientNet-B7 features have been used for multiclass brain tumour classification [3], transfer-learned deep features for diagnosing Down syndrome from facial images [7], and comparative deep-feature analysis for COVID-19 detection from chest radiographs [5]. Hybrid approaches combine deep and statistical features [8]–[10], and stacked ensembles have been evaluated for skin cancer and COVID-19 detection [11]. Convolutional models have also been validated for glaucoma assessment from fundus images [2] and applied to colon and lung cancer histopathology [4].

On the explainability side, surveys describe the importance, stages, and output forms of XAI [12], [13], and medical imaging reviews highlight challenges of validating explanations clinically [1], [14]. Post-hoc methods such as Grad-CAM [15], LIME [16], and SHAP [17] are widely used, whereas inherently interpretable models are advocated for high-stakes settings [18]. Rule-based approaches have been applied to COVID-19 chest X-ray classification [19], and occlusion sensitivity has been combined with a multi-

dataset network for lung disease classification [20]. Metrics for evaluating saliency maps have been reviewed [21].

Gap addressed. The surveyed literature indicates that many classification studies report performance without accompanying explanation analysis, and that comparative benchmarking of heavy versus lightweight backbones together with class-aware error analysis is less common. This paper contributes a controlled-comparison benchmark and a defined path to explanation analysis; it does not claim to be the first such study.

III. DATASET DESCRIPTION

The supplied material specifies the numbers of images per partition and per class, but it does not identify the dataset, imaging modality, anatomical site, disease, class semantics, acquisition protocol, image resolution, or patient counts. Classes are denoted only as Class 0 and Class 1. Each of these items is not specified in the supplied experimental material and is listed under “Information Required for Final Reproducibility” (Section XIX). Consequently, no clinical meaning (for example, which class corresponds to disease) is assigned to either label in this paper.

The partitions comprise 4,708 training, 524 validation, and 624 test images, totalling 5,856 images (Table I). The method used to create the partitions, and whether splitting was performed at patient or image level, is not specified in the supplied experimental material.

IV. DATASET DISTRIBUTION AND CLASS IMBALANCE

Table I and Fig. 1 summarise the class distribution. The training partition contains 1,214 Class 0 images (25.79%) and 3,494 Class 1 images (74.21%); the validation partition mirrors this ratio (135 vs. 389, 25.76% vs. 74.24%). The test partition differs: 234 Class 0 images (37.50%) and 390 Class 1 images (62.50%).

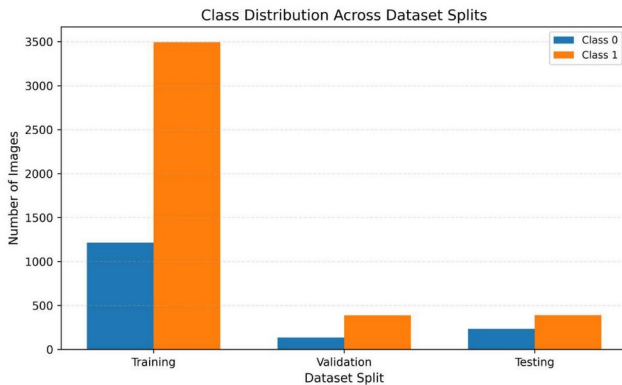


Fig. 1. Class distribution across the training, validation, and testing partitions (supplied figure). Class 1 is the majority class in all partitions; the test partition contains a larger proportion of Class 0 than the training and validation partitions.

TABLE I
DATASET DISTRIBUTION

Dataset	Class 0	Class 1	Total	Class 0 (%)	Class 1 (%)
Training	1214	3494	4708	25.79	74.21
Validation	135	389	524	25.76	74.24
Testing	234	390	624	37.50	62.50

Fig. 1 demonstrates a roughly three-to-one majority of Class 1 in the training and validation data, together with a shift toward Class 0 in the test data. This has two consequences for interpretation. First, a model trained on the skewed distribution may favour the majority class, so recall for Class 0 deserves particular attention. Second, because the test class mix differs from the training mix, test-set accuracy is a support-weighted summary specific to this partition and would change if the class proportions changed. How class imbalance was handled during training (for example, class weighting or resampling) is not specified in the supplied experimental material.

V. DATA PREPROCESSING

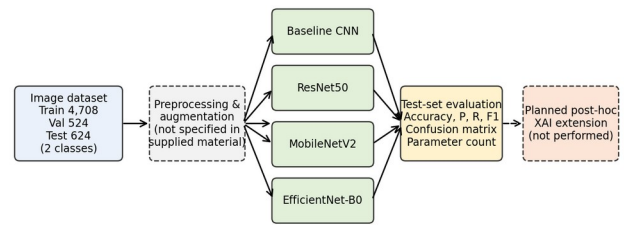
Image resizing, intensity normalisation, colour-space handling, cropping, and any quality filtering were not specified in the supplied experimental material. Upstream choices of this type can alter the information available to a classifier [22], so their documentation is essential for reproducibility.

VI. DATA AUGMENTATION

Whether data augmentation was used, and with which transformations or library [23], is not specified in the supplied experimental material. No augmentation claims are therefore made. Given the imbalance in Fig. 1, augmentation could affect the minority class disproportionately, which is a further reason to report it.

VII. PROPOSED EXPERIMENTAL FRAMEWORK

The experimental framework (Fig. 2) compares one model trained as a baseline with three widely used architectures on the same partitions and evaluates them on the held-out test set. The framework ends with a post-hoc explainability stage that is defined in Section XVIII but was not executed in the supplied experiments.



Solid boxes: supplied experimental results. Dashed boxes: information not supplied / future work.

Fig. 2. Schematic of the experimental framework. Solid boxes correspond to supplied results; dashed boxes denote information not supplied or work not yet performed. The diagram contains no experimental data.

VIII. BASELINE CNN

The baseline CNN contains 422,530 trainable parameters and no non-trainable parameters (Table VI). Its layer configuration (number of convolutional blocks, filter sizes, pooling, regularisation, and classifier head) is not specified in the supplied experimental material. The supplied training curves cover 30 epochs (Fig. 3). Training accuracy rises from roughly 87% in the first epoch to roughly 96% by the final epoch, and training loss falls from approximately 0.32 to approximately 0.11 (values read from the plotted curves; the underlying numeric logs were not supplied). Validation accuracy is erratic in the early epochs, with pronounced dips (for example, to roughly 58% in epoch 1 and below 80% in epochs 3 and 10), and validation loss shows several spikes, the largest near epoch 17 at roughly 0.48. After approximately epoch 20 the validation curves lie close to the training curves, with validation accuracy of roughly 95–96% and validation loss of roughly 0.12–0.18 in the final epochs. The loss axis is labelled cross-entropy.

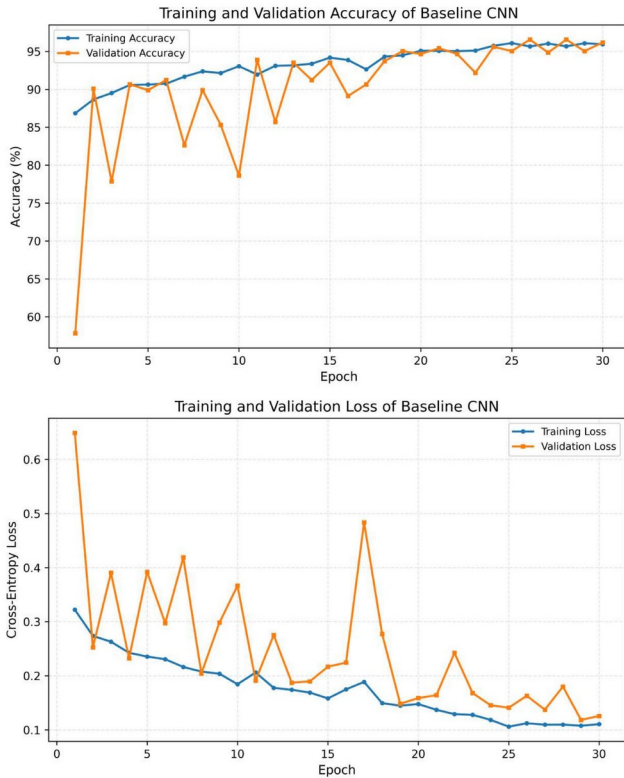


Fig. 3. Training and validation accuracy (top) and cross-entropy loss (bottom) of the baseline CNN over 30 epochs (supplied figures). Validation curves fluctuate substantially in early epochs and approach the training curves in later epochs.

The unstable early validation behaviour is consistent with the small number of validation images and with the class imbalance in the validation set, but the supplied material does not permit the cause to be established. Learning curves for ResNet50, MobileNetV2, and EfficientNet-B0 were not supplied. Note also that the final-epoch validation accuracy (roughly 96%) exceeds the baseline's test accuracy (90.87%); the test partition has a different class mix, and this difference should be borne in mind when comparing validation and test results.

IX. TRANSFER-LEARNING MODELS

A. ResNet50

ResNet50 is a 50-layer residual network whose identity shortcut connections ease the optimisation of deep architectures [6]. In this study it contains 23,512,130 parameters, all reported as trainable (Table VI), which indicates that the whole network was left trainable, although the fine-tuning schedule is not specified in the supplied experimental material.

B. MobileNetV2

MobileNetV2 is a lightweight convolutional architecture designed for efficient computation [REF. TO BE ADDED: original MobileNetV2 paper; not in the approved reference pool]. Here it contains 2,226,434 parameters, all reported as trainable.

C. EfficientNet-B0

EfficientNet-B0 is the smallest member of the EfficientNet family; larger members, such as EfficientNet-B7, have been fine-tuned for medical tasks including multiclass brain tumour classification [3] [REF. TO BE ADDED: original EfficientNet paper; not in the approved reference pool]. Here it contains 4,010,110 parameters, all reported as trainable.

Whether the three architectures were initialised from pre-trained weights, the source of those weights, and the layers that were fine-tuned are not specified in the supplied experimental material. The category “transfer-learning models” is used here because the study design pairs them against a baseline CNN; readers should treat the initialisation detail as an open reproducibility item.

X. TRAINING AND EVALUATION PROTOCOL

The supplied material provides the partition sizes (Table I), the 30-epoch baseline curves (Fig. 3), the test-set metrics, confusion matrices, and parameter counts. The optimiser, learning rate and schedule, batch size, number of epochs for the transfer-learning models, loss function for the transfer-learning models, early-stopping or checkpoint-selection rule, random seeds, hardware, and software versions are not specified in the supplied experimental material. All reported performance values are computed on the 624-image test partition from a single evaluation per model; no repeated runs or cross-validation were supplied.

XI. PERFORMANCE METRICS

Let TP, TN, FP, and FN denote the entries of the confusion matrix for a class treated as positive. The following standard definitions are used. These metrics are conventional, and no additional reference is cited for them.

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) \quad (1)$$

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \quad (2)$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) \quad (3)$$

$$\text{F1} = 2 \cdot \text{Precision} \cdot \text{Recall} / (\text{Precision} + \text{Recall}) \quad (4)$$

Accuracy is the fraction of correctly classified images. **Precision** is the fraction of images predicted as a class that truly belong to it. **Recall** is the fraction of images of a class that are correctly identified. **F1-score** is the harmonic mean of precision and recall. **Weighted** averages combine class-wise values using class support as weights (234 for Class 0 and 390 for Class 1), so the weighted recall equals overall accuracy, as visible in Table II. The **confusion matrix** tabulates predicted against actual classes and shows the direction of every error. Rows denote actual classes and columns denote predicted classes throughout.

XII. EXPERIMENTAL RESULTS

A. Overall performance

Table II and Figs. 4 and 5 report the overall results on the test set. MobileNetV2 attained the highest accuracy (95.03%) and weighted F1-score (94.97%), followed by EfficientNet-B0 (94.23%, 94.13%), ResNet50 (93.43%, 93.31%), and the baseline CNN (90.87%, 90.70%). Within each model, weighted precision is slightly higher than weighted recall, which is consistent with precision benefiting from the very high Class 0 precision values in Table III.

TABLE II
OVERALL MODEL PERFORMANCE ON THE TEST SET
(WEIGHTED AVERAGES)

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Baseline CNN	90.87	91.11	90.87	90.70
ResNet50	93.43	93.85	93.43	93.31
MobileNetV2	95.03	95.24	95.03	94.97
EfficientNet-B0	94.23	94.59	94.23	94.13

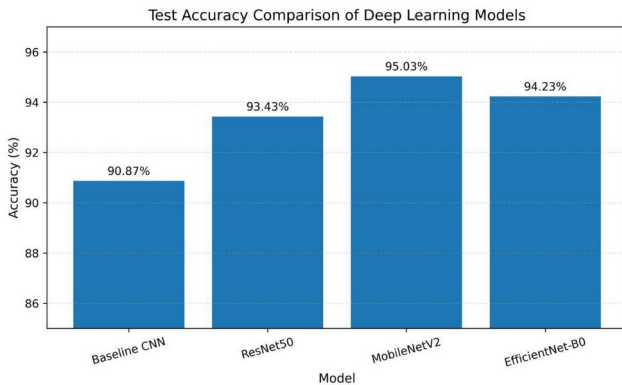


Fig. 4. Test accuracy comparison of the baseline CNN, ResNet50, MobileNetV2, and EfficientNet-B0 (supplied figure). The vertical axis is truncated at 85%, which visually exaggerates between-model differences.

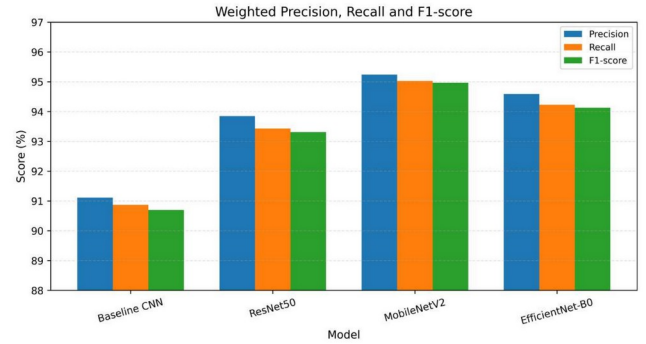


Fig. 5. Weighted precision, recall, and F1-score for the four models (supplied figure). The vertical axis begins at 88%.

Figs. 4 and 5 show the same ranking in all four summary metrics. Because both vertical axes are truncated, bar heights should not be read as proportional differences; the numerical values in Table II are the authoritative representation.

B. Class-wise performance

Table III presents class-wise results. For every model, Class 1 recall (96.92%–99.49%) exceeds Class 0 recall (80.77%–88.03%), whereas Class 0 precision (94.03%–99.01%) exceeds Class 1 precision (89.36%–93.25%). In other words, the models rarely label an image as Class 0 incorrectly, but they miss a notable share of true Class 0 images, assigning them to Class 1. MobileNetV2 achieved the highest Class 0 recall (88.03%) and Class 0 F1-score (93.00%), while EfficientNet-B0 achieved the highest Class 0 precision (99.01%) and Class 1 recall (99.49%).

TABLE III
CLASS-WISE PRECISION, RECALL, F1-SCORE, AND SUPPORT

Model	Class	Precision (%)	Recall (%)	F1-score (%)	Support
Baseline CNN	Class 0	94.03	80.77	86.90	234
Baseline CNN	Class 1	89.36	96.92	92.99	390
ResNet50	Class 0	98.49	83.76	90.53	234
ResNet50	Class 1	91.06	99.23	94.97	390
MobileNetV2	Class 0	98.56	88.03	93.00	234
MobileNetV2	Class 1	93.25	99.23	96.15	390
EfficientNet-B0	Class 0	99.01	85.47	91.74	234
EfficientNet-B0	Class 1	91.94	99.49	95.57	390

C. Confusion matrices

Fig. 6 and Table IV give the confusion matrices. Of the 234 Class 0 test images, the baseline CNN classified 189 correctly and 45 as Class 1; ResNet50, MobileNetV2, and EfficientNet-B0 classified 196, 206, and 200 correctly, respectively. Class 1 errors are far fewer: 12 for the baseline CNN, 3 for ResNet50 and MobileNetV2, and 2 for EfficientNet-B0.

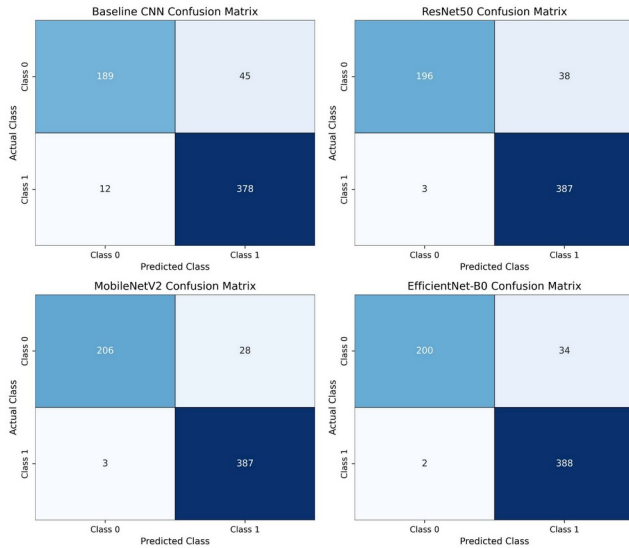


Fig. 6. Confusion matrices on the test set for the baseline CNN, ResNet50, MobileNetV2, and EfficientNet-B0 (supplied figures; rows: actual class, columns: predicted class).

TABLE IV
CONFUSION-MATRIX VALUES (ROWS: ACTUAL; COLUMNS: PREDICTED)

Model	C0 → C0	C0 → C1	C1 → C0	C1 → C1
Baseline CNN	189	45	12	378
ResNet50	196	38	3	387
MobileNetV2	206	28	3	387
EfficientNet-B0	200	34	2	388

Note: C0 = Class 0, C1 = Class 1; “C0 → C1” denotes actual Class 0 images predicted as Class 1.

D. Improvement over the baseline

Table V reports accuracy gains relative to the baseline CNN. Absolute improvement is expressed in percentage points and relative improvement as a percentage of the baseline accuracy (90.87%). MobileNetV2 shows the largest gain (4.16 points; 4.58%), followed by EfficientNet-B0 (3.36; 3.70%) and ResNet50 (2.56; 2.82%). A dedicated bar-chart of these gains was not included in the supplied figure file; the values in Table V are used instead, and no new graph has been drawn.

TABLE V
IMPROVEMENT OVER THE BASELINE CNN

Model	Accuracy (%)	Absolute improvement (percentage points)	Relative improvement (%)
Baseline CNN	90.87	0.00	0.00
ResNet50	93.43	2.56	2.82
MobileNetV2	95.03	4.16	4.58
EfficientNet-B0	94.23	3.36	3.70

E. Model size

Table VI and Fig. 7 compare parameter counts. The baseline CNN is smallest (422,530), followed by MobileNetV2 (2,226,434), EfficientNet-B0 (4,010,110), and ResNet50 (23,512,130). All parameters are reported as trainable for all four models. ResNet50 has about 10.6 times

as many parameters as MobileNetV2 (a derived ratio), yet its test accuracy is 1.60 percentage points lower. Fig. 8 plots accuracy against parameter count and shows that the two most compact pre-trained models lie above ResNet50 on the accuracy axis.

TABLE VI
MODEL PARAMETER COMPARISON

Model	Total	Trainable	Non-trainable
Baseline CNN	422,530	422,530	0
ResNet50	23,512,130	23,512,130	0
MobileNetV2	2,226,434	2,226,434	0
EfficientNet-B0	4,010,110	4,010,110	0

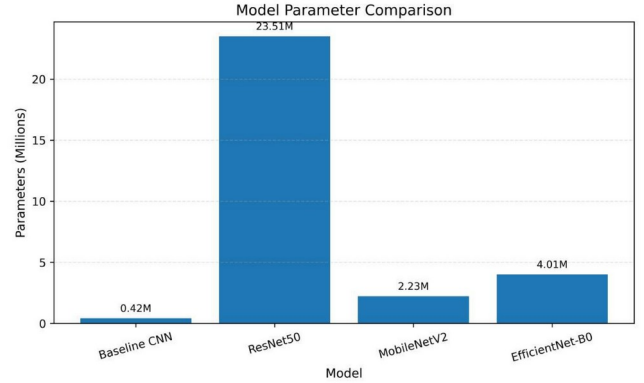


Fig. 7. Parameter counts (in millions) of the four models (supplied figure): 0.42 M, 23.51 M, 2.23 M, and 4.01 M for the baseline CNN, ResNet50, MobileNetV2, and EfficientNet-B0, respectively.

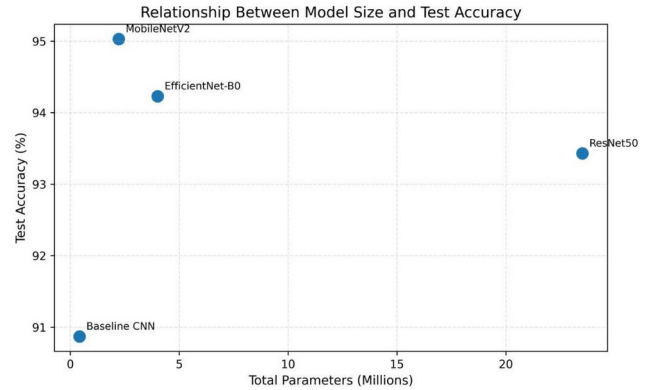


Fig. 8. Relationship between total parameters (millions) and test accuracy (%) for the four models (supplied figure).

XIII. COMPARATIVE ANALYSIS

All three pre-trained architectures outperformed the baseline CNN on every aggregate metric (Table II) and on Class 0 recall (Table III). The baseline’s small size (0.42 M parameters) coincides with the lowest accuracy, but size alone does not determine performance: ResNet50, with the largest parameter count, ranks third. Accuracy therefore does not increase monotonically with model size in this experiment (Fig. 8). Differences between the three pre-trained models are small; for instance, the 0.80-percentage-point gap between MobileNetV2 and EfficientNet-B0 corresponds to five test images (593 versus 588 correct), and the 1.60-point gap to ResNet50 corresponds to ten images (593 versus 583). Because only a single run per

model and no significance test were supplied, these differences should be treated as observed rather than statistically established. Within the evaluated dataset, the ranking MobileNetV2 > EfficientNet-B0 > ResNet50 > baseline CNN is consistent across accuracy, weighted precision, weighted recall, and weighted F1-score.

XIV. ERROR ANALYSIS

Table VII derives error counts from the confusion matrices in Table IV; the numbers are arithmetic consequences of the supplied values and introduce no new measurements. Total misclassifications fall from 57 for the baseline CNN to 41 (ResNet50), 36 (EfficientNet-B0), and 31 (MobileNetV2). In every model, Class 0 images predicted as Class 1 account for the large majority of errors (78.9%–94.4%).

TABLE VII
ERROR COUNTS DERIVED FROM TABLE IV

Model	Total errors	C0 → C1	C1 → C0	C0 → C1 share of errors (%)	Class 0 error rate (%)	Class 1 error rate (%)
Baseline CNN	57	45	12	78.9	19.23	3.08
ResNet50	41	38	3	92.7	16.24	0.77
MobileNetV2	31	28	3	90.3	11.97	0.77
EfficientNet-B0	36	34	2	94.4	14.53	0.51

Two observations follow. First, the gains over the baseline arise from both error types: Class 1 errors fall from 12 to 2–3, and Class 0 errors fall from 45 to 28–38. In absolute counts, ResNet50 removes 7 Class 0 and 9 Class 1 errors, MobileNetV2 removes 17 and 9, and EfficientNet-B0 removes 11 and 10 (arithmetic differences from Table IV). Second, even the best model misclassifies 11.97% of Class 0 test images. Which class is “harder” is therefore Class 0, on the evidence of recall, F1-score, and error rate, although its precision is higher. Possible contributing factors include the 3:1 training imbalance and the test-set shift toward Class 0 (Fig. 1); however, the supplied material does not allow these factors to be separated from intrinsic difficulty, label noise, or image characteristics. Whether errors cluster on particular patients, scanners, or image features is unknown, and no image-level error inspection was supplied.

The clinical meaning of the two error types depends on which class corresponds to the condition of interest, which is not specified in the supplied experimental material. If Class 0 were the diseased class, Class 0 → Class 1 errors would be missed cases; if Class 1 were the diseased class, they would be false alarms. This paper deliberately draws no such conclusion.

XV. DISCUSSION

Best-performing model. Within the evaluated dataset, MobileNetV2 achieved the highest reported accuracy (95.03%), weighted F1-score (94.97%), Class 0 recall (88.03%), and Class 0 F1-score (93.00%). EfficientNet-B0 achieved the highest Class 1 recall (99.49%) and Class 0 precision (99.01%).

Why MobileNetV2 matters. MobileNetV2 achieved the best reported accuracy while using about one-tenth of the parameters of ResNet50 (2,226,434 versus 23,512,130; Table VI). The observed performance suggests that, for this task, additional depth and capacity did not translate into higher test accuracy. Compact models are of practical interest because smaller parameter counts generally imply lower memory requirements; however, parameter count is only a proxy, and no inference time, memory, or energy measurements were supplied. Any statement about deployment suitability would be speculative.

EfficientNet-B0 versus MobileNetV2. EfficientNet-B0 has about 1.8 times as many parameters as MobileNetV2 (derived), reaches 0.80 points lower accuracy, and makes fewer Class 1 errors (2 versus 3) but more Class 0 errors (34 versus 28). The models are thus close in aggregate but differ slightly in error balance. Without repeated runs, we cannot say whether the ordering is stable.

Transfer learning versus the baseline CNN. All three pre-trained models outperformed the baseline by 2.56–4.16 percentage points (Table V). Pre-trained representations are commonly credited for this advantage when labelled medical data are limited [3], [5], [7], but the supplied material does not state whether pre-trained weights were used or how the baseline was trained relative to the others, so the attribution of the gain to transfer learning is a hypothesis consistent with the literature, not a conclusion of this study.

What the confusion matrices reveal. They show that the main residual weakness is Class 0 recall, and that models differ chiefly in how many Class 0 images they push into Class 1 (Table IV and Table VII). Aggregate scores alone would hide this asymmetry.

Effect of class imbalance on interpretation. With 74% of training images in Class 1 and a different test mix, overall accuracy and weighted averages are dominated by the majority class and by the particular test composition. A hypothetical test set with a different class balance would yield different weighted scores for the same models. For this reason, class-wise metrics and the confusion matrix are the more informative summaries.

Why accuracy alone is insufficient in medical diagnosis. Different error types usually carry different consequences, and accuracy treats them equally. Precision quantifies the reliability of positive predictions, recall quantifies how many true cases are found, F1-score balances them, and confusion matrices expose direction and count of each error. Together they allow stakeholders to reason about consequences; accuracy alone does not. Reviews of AI in radiology similarly stress that performance reporting must be accompanied by attention to interpretability and clinical context [1], [14].

Relevance of explainability. High accuracy does not reveal whether a model uses diagnostically meaningful evidence or dataset artefacts, and clinicians need ways to scrutinise model behaviour [1], [24]. Post-hoc XAI can support this scrutiny, but explanation methods must

themselves be validated; otherwise they may give a false sense of assurance [18], [21]. We therefore position XAI as a complement to, not a substitute for, rigorous performance evaluation.

Limitations and future XAI experiments are set out in Sections XVII–XIX.

XVI. EXPLAINABILITY AND XAI DISCUSSION

Five concepts are often conflated and are distinguished here. *Model performance* is the predictive quality measured in Section XII. *Interpretability* is the degree to which a human can understand how a model works, while *explainability* refers to information produced about a model's output, usually post hoc [13], [25], [26]. *Faithfulness* is whether an explanation reflects the evidence the model used. *Localisation* is whether highlighted regions coincide with annotated structures [21]. *Clinical usefulness* is whether explanations improve clinical decisions, which must be established with clinicians [27].

The following post-hoc families are relevant to the four models evaluated here. **Feature attribution** assigns importance to input features and may be viewed as measuring the effect of removing them [28], [29]. **LIME** fits a local surrogate model [16], and **SHAP** decomposes a prediction into additive contributions [17]. **Grad-CAM** uses gradients to produce class-discriminative heat maps over convolutional features [15]. **Occlusion/sensitivity analysis** masks regions and observes output change [20]. **Concept-based** methods test sensitivity to human-defined concepts [30]; **example-based** methods retrieve similar cases [31], [32]; **counterfactual** methods explore minimal changes that alter a prediction [33]; and **rule-based** approaches provide if-then explanations [19], [34]. A synthesis of strengths and limitations is given in the companion review.

No XAI outputs were included in the supplied experimental material. Accordingly, this paper reports no heat maps, attribution scores, localisation measures, or faithfulness results, and it makes no claim that any model's decisions are explained.

XVII. LIMITATIONS

- The dataset, modality, disease, and class meanings are not specified, which prevents clinical interpretation and external comparison.
- Preprocessing, augmentation, optimiser, hyperparameters, transfer-learning protocol, hardware, and software are not specified; the study is not currently reproducible.
- Results come from a single evaluation per model on 624 test images; no variance, confidence intervals, or significance tests were supplied, and small accuracy differences correspond to a handful of images.
- The test partition has a class mix different from the training and validation partitions, and the cause of this shift (and the splitting level, patient versus image) is not stated; potential data leakage cannot be excluded or confirmed.

- Only the baseline CNN has supplied learning curves; the training behaviour of the other models cannot be assessed.
- Metrics are limited to accuracy, precision, recall, F1-score, and confusion matrices; calibration, ROC/PR curves, and threshold analysis were not supplied.
- No inference-time, memory, or energy measurements were supplied, so efficiency is characterised only by parameter count.
- No external validation and no explainability evaluation were performed.

XVIII. PLANNED XAI EVALUATION / FUTURE EXPERIMENTAL EXTENSION

The following extension is proposed as future work. None of it has been executed, and no results are reported. Table VIII outlines the components. The design follows the gaps identified in the companion review: pair explanation with benchmarking, evaluate quantitatively, compare across backbones, and consider class imbalance.

TABLE VIII
PLANNED POST-HOC XAI EVALUATION (NOT YET PERFORMED)

Component	Planned analysis	Intended outcome measure
Grad-CAM [15]	Heat maps for correctly and incorrectly classified test images for each model; layer choice documented	Qualitative review; localisation if annotations exist [21]
Occlusion sensitivity [20]	Mask-and-predict maps with documented occluder size and stride	Faithfulness-style deletion tests [21], [28]
LIME / SHAP [16], [17]	Local explanations for a stratified sample of images	Agreement across methods; stability
Class-aware analysis	Separate analysis of Class 0 → Class 1 errors	Comparison of explanations in errors versus correct cases
Concept/example-based [30], [31]	Optional, if clinically defined concepts or case banks are available	Concept sensitivity; retrieved-case relevance
Expert review [14], [27]	Clinician assessment of explanation usefulness	Structured rating protocol

Particular questions of interest are whether explanations differ between heavy and lightweight backbones, whether the models focus on similar regions when they err on Class 0 images, and whether attribution patterns correlate with the observed recall gap. Planned evaluation should apply quantitative saliency metrics [21] and report all settings.

XIX. INFORMATION REQUIRED FOR FINAL REPRODUCIBILITY

The items below should be completed by the authors from their original experiment logs before submission. The value for each is currently “Not specified in the supplied experimental material.”

TABLE IX
INFORMATION REQUIRED FOR FINAL REPRODUCIBILITY

Item	Status
Dataset name, source, licence, modality, anatomical site	Not specified in the supplied experimental material.
Meaning of Class 0 and Class 1	Not specified in the supplied experimental material.
Image resolution; patient counts; acquisition details	Not specified in the supplied experimental material.

Item	Status
Train/validation/test split method; patient-level separation	Not specified in the supplied experimental material.
Preprocessing and normalisation	Not specified in the supplied experimental material.
Augmentation methods and parameters	Not specified in the supplied experimental material.
Imbalance handling (weights, resampling)	Not specified in the supplied experimental material.
Pre-trained weights, source, frozen/fine-tuned layers	Not specified in the supplied experimental material.
Optimiser, learning rate, schedule, batch size	Not specified in the supplied experimental material.
Epochs for ResNet50, MobileNetV2, EfficientNet-B0 (baseline CNN: 30 epochs in supplied curves)	Not specified in the supplied experimental material.
Checkpoint/early-stopping rule; random seeds	Not specified in the supplied experimental material.
Baseline CNN architecture details	Not specified in the supplied experimental material.
Hardware, software, library versions	Not specified in the supplied experimental material.

XX. CONCLUSION AND FUTURE WORK

This paper compared a baseline CNN with ResNet50, MobileNetV2, and EfficientNet-B0 on a two-class medical image classification task with imbalanced training data and a differently balanced test set. Within the evaluated dataset, the pre-trained models outperformed the baseline, with MobileNetV2 attaining the highest reported accuracy (95.03%) at about one-tenth of the ResNet50 parameter count. Class-wise and confusion-matrix analysis showed that all models share a weakness on Class 0, which accounts for most of the remaining errors, and that aggregate metrics conceal this asymmetry. These findings are limited by the unspecified dataset and training protocol, the single-run design, and the absence of explainability results.

Future work will (i) document and release the full experimental protocol; (ii) repeat training across seeds and report variability; (iii) add calibration and threshold analyses; (iv) validate externally; and (v) carry out the planned post-hoc XAI evaluation of Section XVIII, including class-aware and clinician-informed assessment. No clinical claim is made by this study.

REFERENCES

- [1] B. H. van der Velden, H. J. Kuijf, K. G. Gilhuijs, and M. A. Viergever, "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis," *Med. Image Anal.*, vol. 79, Art. no. 102470, 2022.
- [2] A. Diaz-Pinto, S. Morales, V. Naranjo, T. Köhler, J. M. Mossi, and A. Navea, "CNNs for automatic glaucoma assessment using fundus images: An extensive validation," *Biomed. Eng. Online*, vol. 18, pp. 1–19, 2019.
- [3] A. Ghosh, B. Soni, and U. Baruah, "Transfer learning-based deep feature extraction framework using fine-tuned EfficientNet-B7 for multiclass brain tumor classification," *Arab. J. Sci. Eng.*, vol. 49, no. 9, pp. 12027–12048, 2024.
- [4] N. Ullah, I. De Falco, and G. Sannino, "A novel deep learning approach for colon and lung cancer classification using histopathological images," in *Proc. 2023 IEEE 19th Int. Conf. e-Science*, 2023, pp. 1–10.
- [5] N. Ullah and A. Javed, "Deep features comparative analysis for COVID-19 detection from the chest radiograph images," in *Proc. 2021 Int. Conf. Frontiers Inf. Technol.*, 2021, pp. 258–263.
- [6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
- [7] A. Raza, K. Munir, M. S. Almutairi, and R. Sehar, "Novel transfer learning based deep features for diagnosis of Down syndrome in children using facial images," *IEEE Access*, 2024.
- [8] S. Iqbal, A. N. Qureshi, M. Alhussein, K. Aurangzeb, I. A. Choudhry, and M. S. Anwar, "Hybrid deep spatial and statistical feature fusion for accurate MRI brain tumor classification," *Front. Comput. Neurosci.*, vol. 18, Art. no. 1423051, 2024.
- [9] Z. Lai and H. Deng, "Medical image classification based on deep features extracted by deep model and statistic feature fusion with multilayer perceptron," *Comput. Intell. Neurosci.*, vol. 2018, Art. no. 2061516, 2018.
- [10] G. Latif, D. A. Iskandar, J. M. Alghazo, and N. Mohammad, "Enhanced MR image classification using hybrid statistical and wavelets features," *IEEE Access*, vol. 7, pp. 9634–9644, 2018.
- [11] H. Qayyum, S. T. H. Rizvi, M. Naeem, U. B. Khalid, M. Abbas, and A. Coronato, "Enhancing diagnostic accuracy for skin cancer and COVID-19 detection: A comparative study using a stacked ensemble method," *Technologies*, vol. 12, no. 9, p. 142, 2024.
- [12] N. Ullah, J. A. Khan, I. De Falco, and G. Sannino, "Explainable artificial intelligence: Importance, use domains, stages, output shapes, and challenges," *ACM Comput. Surv.*, vol. 57, no. 4, pp. 1–36, 2024.
- [13] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.
- [14] M. Reyes *et al.*, "On the interpretability of artificial intelligence in radiology: Challenges and opportunities," *Radiol. Artif. Intell.*, vol. 2, no. 3, Art. no. e190043, 2020.
- [15] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.
- [16] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in *Proc. KDD*, 2016, pp. 1135–1144.
- [17] S. Lundberg, "A unified approach to interpreting model predictions," arXiv:1705.07874, 2017.
- [18] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nat. Mach. Intell.*, vol. 1, no. 5, pp. 206–215, 2019.
- [19] I. De Falco, G. De Pietro, and G. Sannino, "Classification of Covid-19 chest X-ray images by means of an interpretable evolutionary rule-based approach," *Neural Comput. Appl.*, vol. 35, no. 22, pp. 16061–16071, 2023.
- [20] N. Ullah, J. A. Khan, I. De Falco, and G. Sannino, "Bridging clinical gaps: Multi-dataset integration for reliable multi-class lung disease classification with DeepCRINet and occlusion sensitivity," in *Proc. 2024 IEEE Symp. Comput. Commun.*, 2024, pp. 1–6.
- [21] T. Gomez, T. Fréour, and H. Mouchère, "Metrics for saliency map evaluation of deep learning explanation methods," in *Int. Conf. Pattern Recognit. Artif. Intell.*, 2022, pp. 84–95.
- [22] V. Bevilacqua, A. Di Marino, E. Di Nardo, A. Ciaramella, I. De Falco, and G. Sannino, "Cross-domain super-resolution in medical imaging," in *Proc. 2024 IEEE Symp. Comput. Commun.*, 2024, pp. 1–6.
- [23] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and flexible image augmentations," *Information*, vol. 11, no. 2, 2020.
- [24] A. I. Poon and J. J. Sung, "Opening the black box of AI-medicine," *J. Gastroenterol. Hepatol.*, vol. 36, no. 3, pp. 581–584, 2021.
- [25] N. Burkart and M. F. Huber, "A survey on the explainability of supervised machine learning," *J. Artif. Intell. Res.*, vol. 70, pp. 245–317, 2021.
- [26] M. Christoph, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Leanpub, 2020.

- [27] G. Stiglic, P. Kocbek, N. Fijacko, M. Zitnik, K. Verbert, and L. Cilar, "Interpretability of machine learning-based prediction models in healthcare," *WIREs Data Mining Knowl. Discov.*, vol. 10, no. 5, Art. no. e1379, 2020.
- [28] I. Covert, S.-I. Lundberg, and S.-I. Lee, "Explaining by removing: A unified framework for model explanation," *J. Mach. Learn. Res.*, vol. 22, no. 209, pp. 1–90, 2021.
- [29] M. Wojtas and K. Chen, "Feature importance ranking for deep learning," in *Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 5105–5114.
- [30] B. Kim *et al.*, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2668–2677.
- [31] N. Papernot and P. McDaniel, "Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning," arXiv:1803.04765, 2018.
- [32] A. Poché, L. Hervier, and M.-C. Bakkay, "Natural example-based explainability: A survey," in *Proc. World Conf. Explainable Artif. Intell.*, 2023, pp. 24–47.
- [33] S. Verma, J. Dickerson, and K. Hines, "Counterfactual explanations for machine learning: A review," arXiv:2010.10596, 2020.
- [34] T. Wang, C. Rudin, Y. Doshi-Velez, Y. Liu, E. Klampfl, and P. MacNeille, "A Bayesian framework for learning rule sets for interpretable classification," *J. Mach. Learn. Res.*, vol. 18, no. 70, pp. 1–37, 2017.