

Two-Stage Transfer Learning for Slice-Level Pancreatic CT Classification: Performance Benchmarking, Confidence-Interval Reporting, Saliency Audit and a Data-Leakage Risk Analysis

Navyasri [surname]

[Affiliation, city, country] · Corresponding author: vemulawada148@gmail.com

Manuscript draft. Items in [square brackets] must be completed by the authors before submission.

Abstract

Background. Pancreatic ductal adenocarcinoma is usually detected late, and computer-aided reading of abdominal CT is an active route to earlier detection. Many published convolutional-network studies report accuracies above 96%, yet most train and evaluate on 2D slices and do not state whether slices from the same patient were kept in the same data partition.

Methods. We trained and compared two ImageNet-pretrained backbones on 880 axial abdominal CT slices (205 normal-labelled, 675 positive-labelled) for binary classification: a frozen MobileNetV2 baseline and a two-stage EfficientNet-B0 pipeline (head training followed by fine-tuning of the final 30 layers with frozen batch-normalisation). Data were split 70/15/15 with stratification. We report test performance with exact Wilson 95% confidence intervals, balanced accuracy and Matthews correlation coefficient, in addition to accuracy. We audit the model with Grad-CAM and with an inspection of file-naming metadata for patient-level overlap.

Results. On the 132-image held-out test set, EfficientNet-B0 reached accuracy 0.992 (95% CI 0.958–0.999), ROC-AUC 0.9997, sensitivity 1.000 (0.963–1.000), specificity 0.968 (0.838–0.994) and MCC 0.979, with a single false positive in 132 slices. The MobileNetV2 baseline reached accuracy 0.871 and ROC-AUC 0.972 but specificity of only 0.452, i.e. its accuracy was only 10.6 points above the 76.5% obtained by labelling every test slice positive. Validation accuracy reached 100% after two epochs with a frozen backbone, and file names show that multiple slices from one case exist in the dataset, so the random slice-level split may allow near-duplicate anatomy to appear in both training and test partitions. Grad-CAM maps from the baseline model concentrated on liver, kidney, spine and soft-tissue regions rather than the pancreas.

Conclusions. Transfer learning with partial fine-tuning separates the two classes almost perfectly on a slice-level split, and fine-tuning EfficientNet-B0 removes the baseline's specificity failure. However, the reported figures should be read as an upper bound on clinical performance until they are reproduced with patient-level partitioning and external data. We provide a reproducible leakage-aware protocol and a reporting checklist for slice-based pancreatic CT classifiers.

Keywords: *pancreatic cancer; computed tomography; deep learning; transfer learning; EfficientNet; Grad-CAM; data leakage; medical image classification*

1. Introduction

Pancreatic cancer has one of the poorest prognoses among solid tumours, largely because it is diagnosed at an advanced stage. Detecting subtle pancreatic lesions on routine CT is difficult even for experienced readers, which has motivated a large literature on deep-learning assistance. The strongest evidence to date

comes from large, multicentre, patient-level studies: the PANDA system, trained on non-contrast CT, reported 94.9% sensitivity and 100% specificity on an internal test set of 291 patients and 93.3% sensitivity and 98.8% specificity across nine external centres (5,337 patients) [1].

A second body of work, mostly using small public image collections, reports near-perfect slice-level accuracy with elaborate pipelines combining segmentation, hand-crafted pre-processing, feature selection and metaheuristic hyper-parameter search [2,3]. These studies are valuable for method development, but in several of them the partitioning strategy is not documented, confidence intervals are absent and internal inconsistencies between tables make the headline numbers hard to interpret. It is well established in other imaging domains that splitting at the image rather than the subject level inflates test performance, by 5 to 30 accuracy points in one systematic OCT study [4].

This paper makes four contributions on a public slice-level pancreatic CT collection:

- A simple, fully specified two-stage EfficientNet-B0 transfer-learning recipe, benchmarked against a frozen MobileNetV2 baseline, with complete training trajectories.
- Uncertainty-aware reporting: exact confidence intervals, balanced accuracy, MCC and a majority-class reference, which expose a weakness (specificity 0.452) that headline accuracy of the baseline hides.
- A saliency plausibility audit with Grad-CAM, reported without selecting favourable examples.
- An explicit data-leakage risk analysis for slice-level datasets, with a patient-grouped protocol (Appendix A) that others can apply to the same data.

2. Related Work

Table 1 summarises the studies most closely related to ours. We list what each study reports and what it does not state, because the missing items (partition unit, confidence intervals, external validation) determine how far the numbers can be trusted.

Table 1. Comparison with representative published studies.

Study	Data	Approach	Reported performance	Not stated / limitation
Cao et al., Nat Med 2023 (PANDA) [1]	3,208 train; 291 internal test; 5,337 external (9 centres); 20,530 real-world; non-contrast CT	Multi-stage 3D deep learning for lesion detection	Internal: AUC 0.996, sens 94.9%, spec 100%. External: sens 93.3%, spec 98.8%. Real-world: sens 92.9%, spec 99.9%	Retrospective; gold-standard design with patient-level, multicentre evaluation
Dodda & Muneeswari, PeerJ CS 2025 [2]	Kaggle slice-level CT, 10,870 images; 9,035 train / 1,835 test	Pre-processing + enhanced U-Net + DenseNet-169 features + modified ResNeXt, metaheuristic tuning	Acc 99.85%, sens 99.76%, spec 99.72%; AUC 0.974	Random split of slice-level data; patient-level separation not stated; single dataset; internal naming inconsistencies
Thiruvengadam et al., arXiv 2025 [3]	Public Kaggle "Pancreatic CT Images" (normal vs tumour); size not reported	MAGRes-UNet + DenseNet-121 + HHO/BA selection + ViT/EfficientNet-B3 hybrid	Acc 96.32%, sens 93.3–95.1%, spec 94.8–96.3% (tables differ)	No AUC, no CI, no split description, no external validation
This work	880 CT slices (205 normal / 675 positive); 132 test slices	Frozen MobileNetV2 baseline; two-stage fine-tuned EfficientNet-B0	Acc 0.992 (0.958–0.999), AUC 0.9997, sens 1.000, spec 0.968, MCC 0.979	Slice-level random split; possible patient overlap; single dataset; no external test

Two observations follow. First, performance on slice-level public data (96–99.9%) is similar to or higher than PANDA's patient-level multicentre figures, even though the latter solves a harder, real-world problem; this gap is more plausibly explained by dataset difficulty and partitioning than by model superiority. Second, the size of that possible inflation is not hypothetical. Tampu et al. showed on four OCT datasets that image-level splits raised test accuracy relative to subject-level splits by 4.3 to 29.6 points (Figure 12) [4].

3. Materials and Methods

3.1 Dataset

The dataset consists of 880 axial abdominal CT slices stored as JPEG files in two folders: NPancreas (205 images; class 0, normal-labelled) and PPancreas (675 images; class 1, positive-labelled). Class prevalence is therefore 76.7% positive. File names encode a case identifier and a slice index (e.g. N.CT1202208230494-12, -13, -14; P.CT4202108030081-19 to -23), showing that several consecutive slices were drawn from the same case. [Authors to add: dataset source and licence, number of unique patients per class, scanner/protocol, contrast phase, labelling method and ethics statement.] Representative slices are shown in Figure 1.

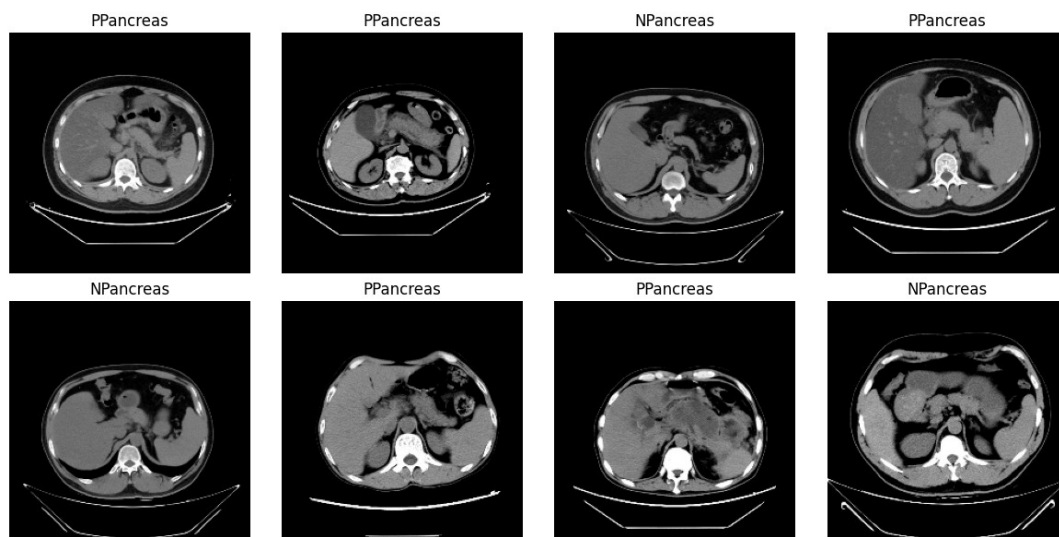


Figure 1. Sample training slices with their folder labels. Windowing, field of view and table geometry vary visibly between slices.

3.2 Data partitioning

Images were split once with stratification and a fixed seed (42) into training (616; 143 normal / 473 positive), validation (132; 31 / 101) and test (132; 31 / 101) sets (Figure 2). The unit of splitting was the image file, not the patient (see Section 5.2).

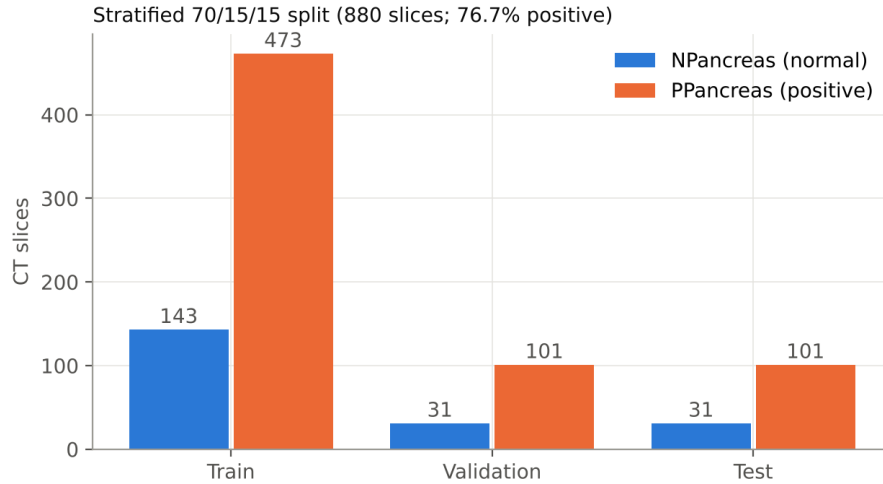


Figure 2. Class composition of the training, validation and test partitions.

3.3 Models and training

All images were resized to 224×224 and converted to three channels. Two pipelines were trained (Figure 3).

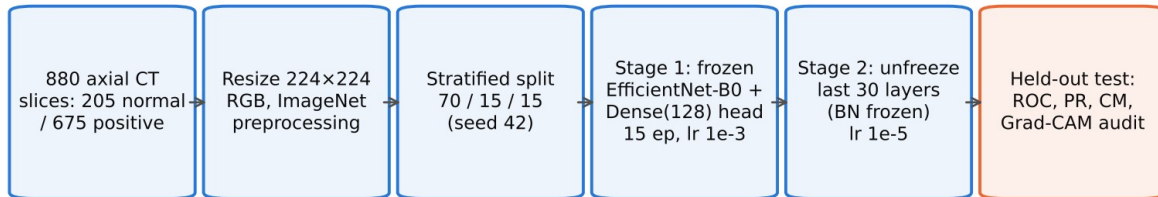


Figure 3. Study pipeline. The EfficientNet-B0 model is trained in two stages; all reported numbers use the final held-out test split.

- Baseline: ImageNet-pretrained MobileNetV2 (frozen) with global average pooling, dropout 0.3 and a one-unit sigmoid output; Adam, learning rate 1e-4, batch size 16, 20 epochs, binary cross-entropy.
- Proposed: ImageNet-pretrained EfficientNet-B0 with global average pooling, dropout 0.35, a 128-unit ReLU layer, dropout 0.25 and a sigmoid output (4,213,668 parameters, 164,097 trainable in stage 1). Stage 1 trained the head for up to 15 epochs (Adam, 1e-3). Stage 2 unfroze the last 30 backbone layers, kept all batch-normalisation layers frozen, and continued for up to 10 epochs at 1e-5.

Both stages used early stopping on validation loss (patience 5, best weights restored), learning-rate reduction on plateau (factor 0.3, patience 2) and checkpointing of the best validation-loss model. Stage 1 ran all 15 epochs; stage 2 stopped after 6 epochs and restored the weights from its second epoch (validation loss 0.0021). No data augmentation was used. Implementation: TensorFlow/Keras, scikit-learn for metrics.

3.4 Evaluation

The test set was used once, after training. We report accuracy, sensitivity (recall for the positive class), specificity, precision, F1, balanced accuracy, MCC, ROC-AUC and the precision–recall curve at a decision threshold of 0.5. Binomial proportions are given with exact Wilson 95% confidence intervals computed from the confusion-matrix counts. Because 76.5% of test slices are positive, we also report the majority-class accuracy as a reference.

3.5 Saliency audit

Grad-CAM [8] was computed on the last convolutional block (out_relu) of the MobileNetV2 baseline for one slice from each folder, chosen as the first file in sorted order rather than for visual quality. The overlay thresholding and the bounding shapes in the figure are illustrative visualisation choices, not tumour segmentations.

4. Results

4.1 Training behaviour

The frozen MobileNetV2 baseline improved slowly, from 71.6% training accuracy in epoch 1 to 86.2% in epoch 20, with validation accuracy flat at 75.8% for the first eight epochs, which equals the validation prevalence of the positive class, i.e. the model was predicting almost everything as positive (validation recall 0.99). EfficientNet-B0 reached 98.5% validation accuracy after one epoch of head training and 100% after two, with validation loss falling from 0.156 to 0.0021 by the end of fine-tuning (Figure 4).

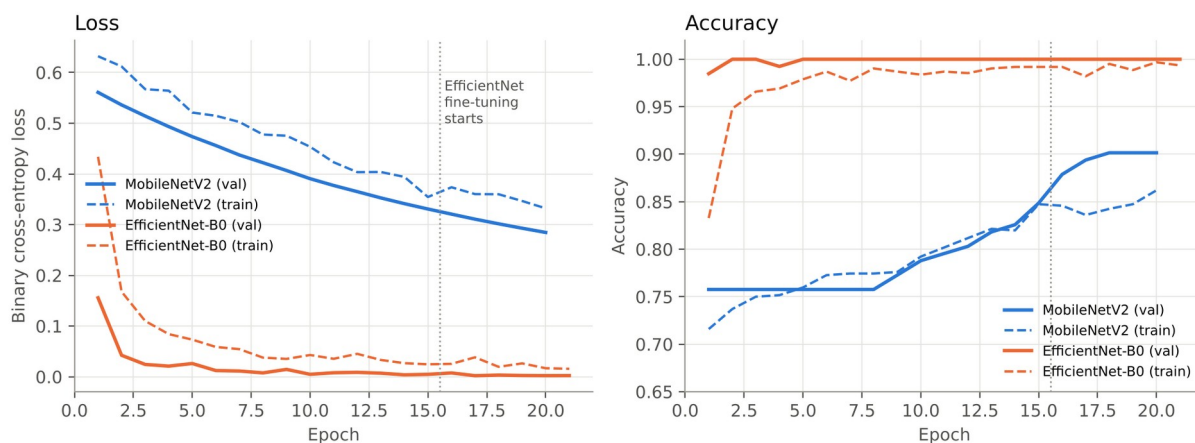


Figure 4. Training (dashed) and validation (solid) loss and accuracy. The dotted line marks the start of EfficientNet-B0 fine-tuning (stage 2).

4.2 Test performance

Table 2. Held-out test performance (n = 132; 31 normal, 101 positive). Majority-class (all-positive) accuracy = 0.765.

Metric	MobileNetV2 (frozen)	EfficientNet-B0 (two-stage)
Accuracy (95% CI)	0.871 (0.803–0.918)	0.992 (0.958–0.999)
Sensitivity (95% CI)	1.000 (0.963–1.000)	1.000 (0.963–1.000)
Specificity (95% CI)	0.452 (0.292–0.622)	0.968 (0.838–0.994)
Precision (95% CI)	0.856	0.990 (0.947–0.998)

F1 score	0.922	0.995
Balanced accuracy	0.726	0.984
Matthews correlation coefficient	0.622	0.979
ROC-AUC	0.972	0.9997
Test loss	0.296	0.021
Confusion matrix (TN / FP / FN / TP)	14 / 17 / 0 / 101	30 / 1 / 0 / 101

The baseline detected every positive slice but misclassified 17 of 31 normal slices (specificity 0.452); its accuracy of 0.871 is only 10.6 points above the all-positive reference. Fine-tuned EfficientNet-B0 produced one error in 132 slices, a normal-labelled slice scored 0.866 for the positive class (Figure 10). Because the test set contains only 31 normal slices, the specificity estimate is imprecise: the lower bound of its 95% interval is 0.838 (Figure 9).

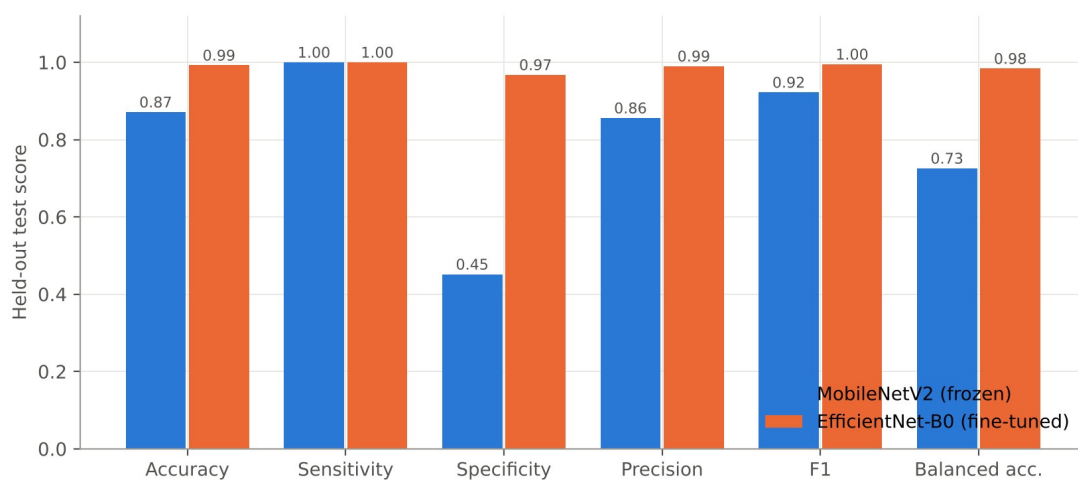


Figure 5. Test-set comparison of the two models across six metrics.

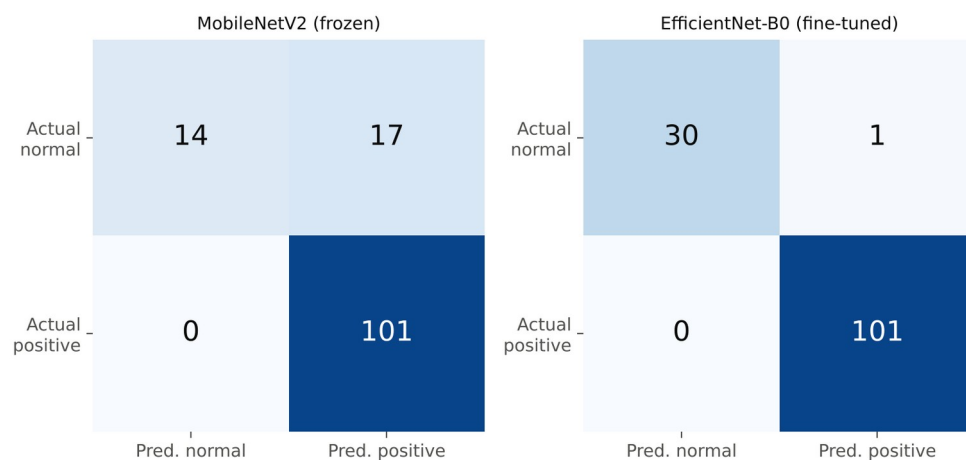


Figure 6. Confusion matrices on the 132-slice test set.

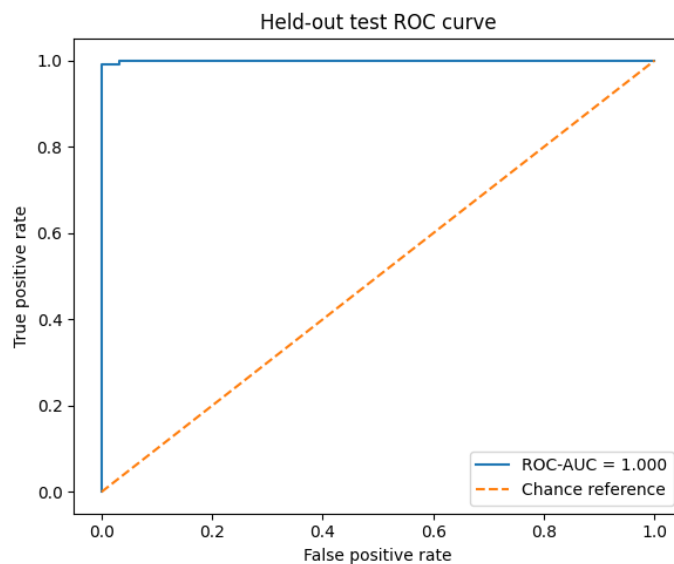


Figure 7. ROC curve of the EfficientNet-B0 model on the test set (AUC 0.9997).

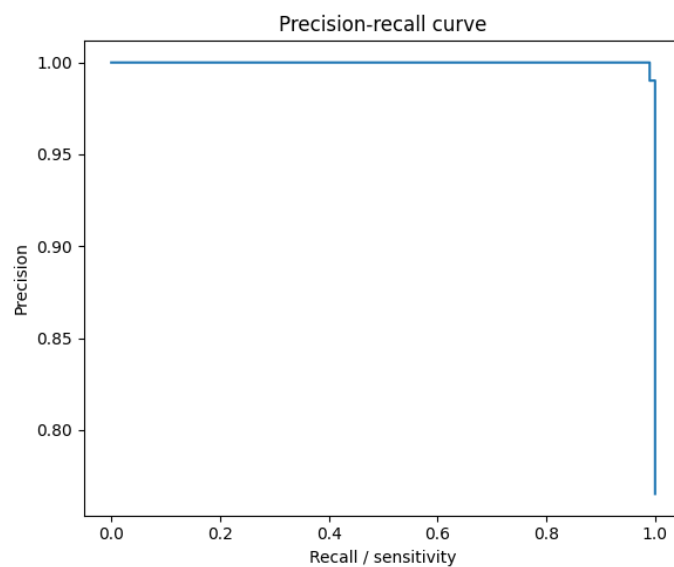


Figure 8. Precision–recall curve of the EfficientNet-B0 model on the test set.

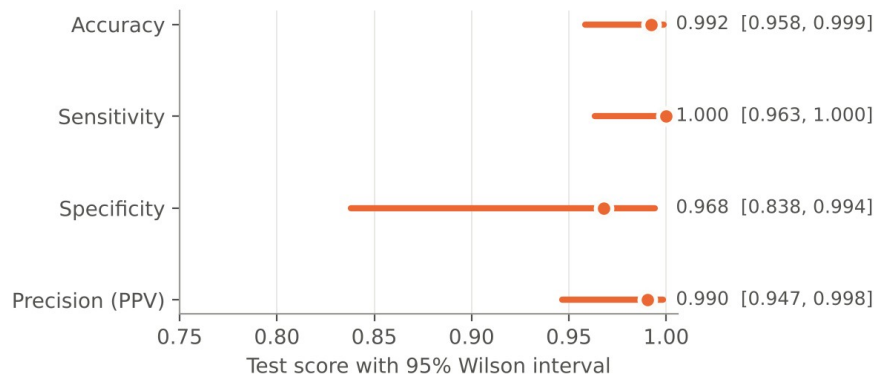


Figure 9. Test metrics of EfficientNet-B0 with exact 95% Wilson confidence intervals. The wide specificity interval reflects only 31 normal test slices.

Normal-labelled Images Predicted as Class 1

Actual: 0
 Predicted: 1
 Class 1 score: 86.6%



Figure 10. The only misclassified test slice: normal-labelled, scored 86.6% for the positive class. There were no false negatives.

4.3 Saliency audit

For the baseline model, Grad-CAM attention for both the normal example (score 0.732) and the positive example (score 0.886) covered large parts of the liver, kidney, spine and surrounding soft tissue rather than a compact pancreatic region (Figure 11). The model assigned a score above 0.5 to the normal example, consistent with the baseline's false-positive rate. These maps do not demonstrate that the classifier uses pancreatic tissue; they are compatible with the model relying on global appearance cues (patient habitus, windowing, contrast phase) that differ between the two folders. Grad-CAM was not run on the final EfficientNet-B0 model in the present notebook; this is the most important pending analysis.

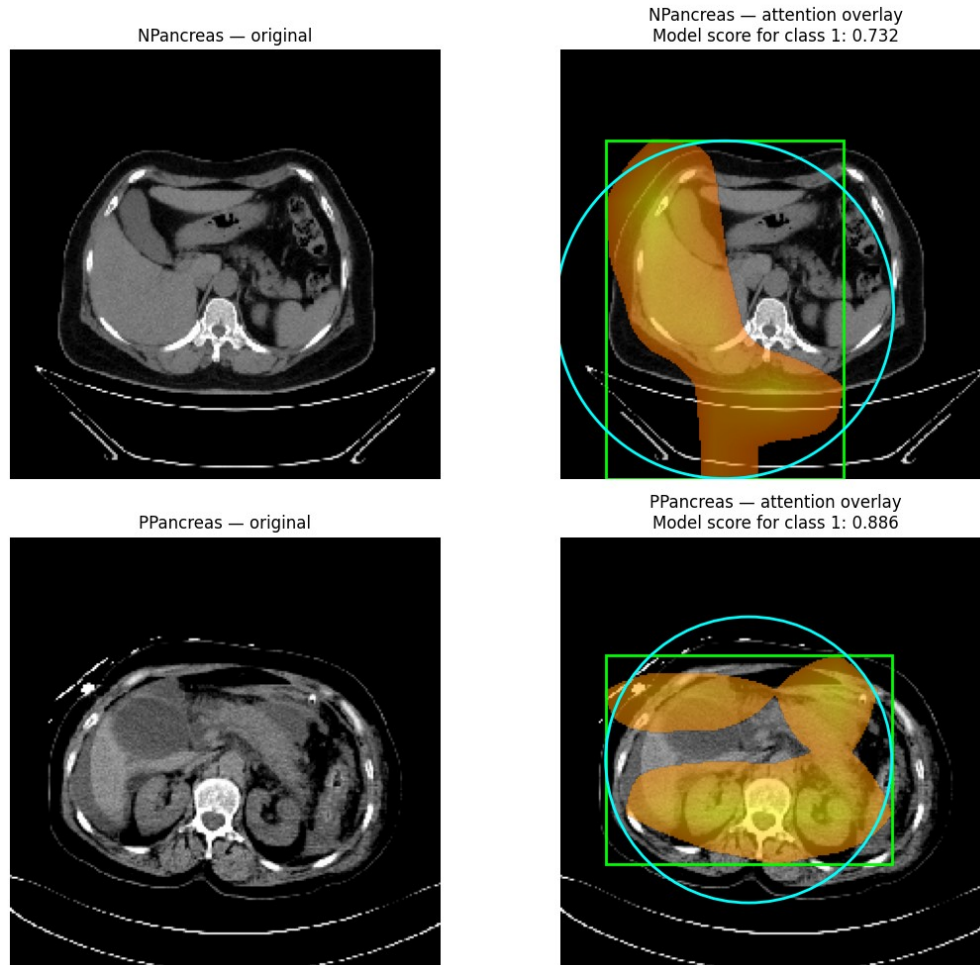


Figure 11. Grad-CAM audit of the MobileNetV2 baseline on the first file of each folder. Top: normal-labelled; bottom: positive-labelled. The coloured overlay marks regions with high class activation; boxes and circles are illustrative annotations, not lesion boundaries. Scores are classifier outputs, not clinical probabilities.

5. Discussion

5.1 Principal findings

Fine-tuning the last 30 layers of EfficientNet-B0, with batch-normalisation frozen, plus a small two-layer head gave near-perfect separation on the held-out split, and the improvement over the frozen MobileNetV2 baseline is concentrated in specificity (0.452 to 0.968) rather than sensitivity (1.000 in both). This pattern shows why accuracy alone is a poor summary on an imbalanced dataset: the baseline's 0.871 accuracy concealed a failure to recognise most normal slices. Reporting sensitivity, specificity, balanced accuracy and MCC with confidence intervals is therefore recommended for all slice-level pancreatic studies.

5.2 Why the numbers should be treated as an upper bound

Several observations indicate that the 99.2% accuracy likely overstates performance on new patients:

- Unit of splitting. The split was random at the image level, and the file names show consecutive slices from the same case. Adjacent CT slices are highly correlated, so a test slice can have near-duplicates in the training set.
- Speed of convergence. With the backbone frozen, a single dense head reached 98.5% validation accuracy after one epoch and 100% after two, which is unusually fast for a task as subtle as pancreatic lesion detection and is typical of leakage or an easy confounder.
- Possible confounding. Image windowing, field of view and table geometry vary (Figure 1), and case identifiers carry different prefixes. If normal and positive cases came from different scanners, dates or protocols, a network can learn those differences instead of pathology.
- Saliency. The baseline's attention was not pancreas-specific (Figure 11).
- Magnitude in the literature. Image-level versus subject-level splitting changed accuracy by 4.3 to 29.6 points in four OCT datasets (Figure 12) [4]. We cannot say how large the effect is for this dataset without the patient-level re-run described in Appendix A.

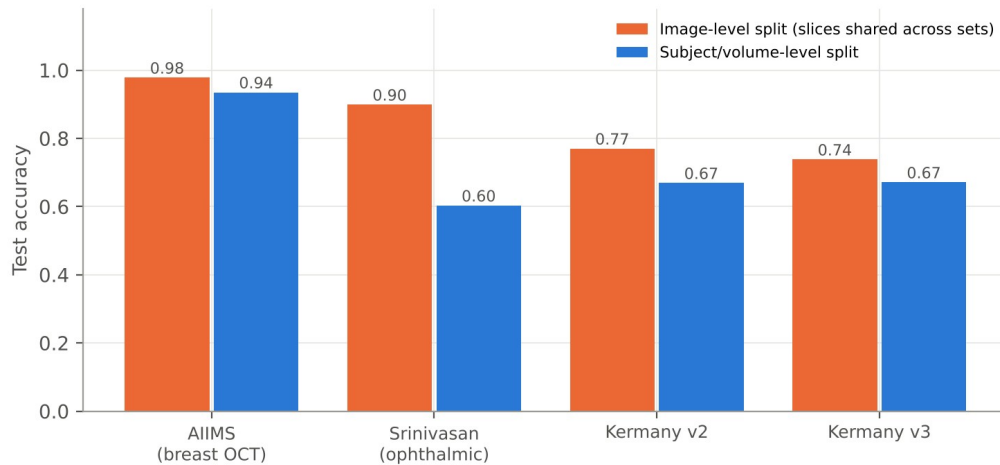


Figure 12. Test accuracy under image-level versus subject-level splitting in four OCT datasets (LightOCT model; values from Tampu et al. [4]). Included to illustrate the size of the leakage effect in a comparable slice-based setting; these are not results from our data.

5.3 Comparison with prior work

Our slice-level accuracy (0.992) lies between that of a pipeline with segmentation, metaheuristic feature selection and a ViT-EfficientNet hybrid (96.3%) [3] and a segmentation-plus-ResNeXt pipeline (99.85%) [2]. The comparison is indicative only: the datasets, partitions and metrics differ, and none of these studies, including ours, reports patient-level validation on this class of public slice data. The point of our result is that a lightweight, fully specified recipe with no segmentation, no augmentation and no metaheuristic search already reaches this accuracy band. This suggests that the headline accuracy on this kind of data discriminates poorly between methods and that complexity is not what the benchmark rewards. PANDA [1] remains the reference for what clinically meaningful validation looks like: patient-level, multicentre, prospective-style evaluation, with sensitivity of 92.9–94.9% and specificity of 98.8–100%.

5.4 Clinical relevance

The model is a research classifier of single 2D slices. It does not localise lesions, does not aggregate over a patient's volume, and has not been evaluated on external data. It must not be used for diagnosis. A clinically useful system would need patient-level aggregation, lesion localisation, calibration and prospective validation.

6. Limitations

- Single public dataset; labels are folder names, with unverified provenance (histology, radiology report or other).
- Image-level split without confirmed patient separation; the number of unique patients is unknown to us.
- Small test set (132 slices, 31 normal): the lower 95% bound for specificity is 0.838, and for accuracy 0.958.
- Single random split and single seed; no cross-validation, no repeated runs, no variance estimate.
- No external or temporal validation; no calibration analysis; fixed 0.5 threshold.
- Grad-CAM performed on the baseline only, on two slices; no quantitative localisation metric.
- The baseline was trained for 20 epochs without early stopping and without fine-tuning, so the model comparison also confounds architecture with training recipe.

7. Conclusion

A two-stage fine-tuned EfficientNet-B0 classifies slices of this pancreatic CT collection with 99.2% test accuracy (95% CI 95.8–99.9%) and removes the specificity failure of a frozen MobileNetV2 baseline. Because the split was at the slice level, the data show multiple slices per case, convergence was extremely fast and saliency was not pancreas-specific, these figures should be regarded as an upper bound pending patient-grouped re-evaluation and external validation. We recommend that studies on slice-level public CT data report the partition unit, confidence intervals, balanced metrics and a majority-class reference as standard.

Declarations

Data availability: [dataset source, DOI/URL and licence]. Code availability: [repository URL]; the notebook used is available from the authors. Ethics: [statement on use of anonymised public data]. Funding: [none / state]. Competing interests: [none / state]. Author contributions: [state]. AI assistance: [state any use of AI tools in drafting, as required by the target journal].

References

- [1] Cao K, Xia Y, Yao J, et al. Large-scale pancreatic cancer detection via non-contrast CT and deep learning. *Nature Medicine* 2023;29(12):3033–3043.
- [2] Dodda R, Muneeswari G. Automatic pancreatic cancer segmentation and classification in CT images using an integrated deep-learning approach. *PeerJ Computer Science* 2025;11:e3263.
- [3] Thiruvengadam JA, Nabigaru KM, Kovi A. Scalable residual feature aggregation framework with hybrid metaheuristic optimization for robust early pancreatic neoplasm detection in multimodal CT imaging. *arXiv:2512.23597*.

- [4] Tampu IE, Eklund A, Haj-Hosseini N. Inflation of test accuracy due to data leakage in deep learning-based classification of OCT images. arXiv:2202.12267, 2022.
- [5] Tan M, Le QV. EfficientNet: rethinking model scaling for convolutional neural networks. Proc ICML 2019.
- [6] Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. MobileNetV2: inverted residuals and linear bottlenecks. Proc CVPR 2018.
- [7] Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. Proc CVPR 2009.
- [8] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. Proc ICCV 2017.
- [9] Kingma DP, Ba J. Adam: a method for stochastic optimization. Proc ICLR 2015.
- [10] Wilson EB. Probable inference, the law of succession, and statistical inference. J Am Stat Assoc 1927;22(158):209–212.

Appendix A. Patient-grouped re-evaluation protocol

To convert the present results into publishable evidence, re-run the notebook with the split below. It derives the case identifier from the file name (the part before the slice index) and keeps every slice of a case in one partition. Report results as mean $\pm$ SD over folds with confidence intervals computed at the case level.

```
import re, numpy as np
from sklearn.model_selection import StratifiedGroupKFold

def case_id(path):
    stem = path.split("/")[-1].rsplit(".", 1)[0]          # N.CT1202208230494-12
    return re.sub(r"-\\d+$", "", stem)                     # N.CT1202208230494

groups = np.array([case_id(p) for p in paths])
print("unique cases:", len(set(groups)))
sgkf = StratifiedGroupKFold(n_splits=5, shuffle=True, random_state=42)
for fold, (tr, te) in enumerate(sgkf.split(paths, labels, groups)):
    assert not set(groups[tr]) & set(groups[te])          # no case in both sets
    # carve a validation set from tr with the same grouped splitter,
    # train the two-stage EfficientNet-B0, evaluate on te
```

Further recommended analyses: (i) Grad-CAM on the final EfficientNet-B0 with a quantitative overlap against pancreas masks if available; (ii) a negative-control test, such as masking the pancreatic region or training on a non-pancreatic crop, to detect confounders; (iii) data augmentation and repeated seeds; (iv) calibration (reliability curve, Brier score) and threshold selection on the validation set; (v) external validation on an independent public or institutional cohort.