COVID-19 Detection in Chest CT Scans: Custom CNN, ResNet18 & Vision Transformer with Explainability via Grad-CAM & MoRF/AOPC

Iraklis Spyrou
University of Pireaus-NCSR Demokritos
MSc "Artificial Intelligence"
Email: iraklis.spyrou@gmail.com

***Abstract*—We benchmark three architectures—a lightweight CustomCNN trained from scratch, an ImageNet-pretrained ResNet18, and a pretrained Vision Transformer (ViT-B/16)—on binary COVID vs. non-COVID classification using a balanced chest-CT dataset of 2,482 slices (1,252 COVID, 1,230 non-COVID). ResNet18 achieves the best performance, reaching 98.9 % accuracy and a macro F1-score of 0.989. To interpret model decisions, we generate SmoothGradCAM++ heatmaps and assess their fidelity with MoRF and AOPC curves. The visual and quantitative analyses show that all transfer-learned models rely on clinically relevant ground-glass opacities for COVID predictions, whereas CustomCNN learns only positive cues and ignores non-COVID evidence. These findings highlight the advantage of transfer learning for both performance and explainability on limited medical-imaging data.**

***Index Terms*—COVID-19, Chest CT, Deep Learning, CNN, ResNet18, Vision Transformer, Grad-CAM, MoRF, AOPC**

I. Introduction

Ever since COVID-19 emerged, healthcare systems everywhere have been under pressure to find faster and more reliable ways to diagnose it. While PCR tests remain the gold standard—pinpointing viral RNA with high accuracy—they rely on lab facilities and can take hours to return results. Chest CT scans, on the other hand, are another reliable medical examination of COVID-19, revealing patterns in the lungs, such as ground glass opacities [1]. Deep learning promises to make CT-based screening even smoother. Models like convolutional neural networks (CNNs) and Vision Transformers (ViTs) can learn to spot subtle patterns in lung images – sometimes matching or even exceeding radiologist-level accuracy. This means quicker triage and less stress on overworked imaging departments [2]. However, the "black-box" nature of these models raises concerns: without insights into their decision-making, it is difficult to ensure reliability and detect potential biases. For example, Grad-CAM is an explainability technique that generates color-coded heatmaps for each CT image, highlighting which pixels most influence the model's prediction. To evaluate these heatmaps quantitatively, we apply MoRF (Most Relevant First), which gradually masks the top-ranked pixels, and AOPC (Area Over the Perturbation Curve), which measures how the model's confidence drops as these important pixels are removed [3]. In this study, we implement and compare three distinct architectures: a custom CNN trained from scratch, a pre-trained ResNet18 and a pretrained ViT-B/16. We evaluate their diagnostic performance on a balanced public dataset of COVID and non-COVID scans. To interpret model decisions, we generate SmoothGradCAM++ heat maps and quantitatively assess them using MoRF and AOPC metrics, thereby offering both accurate detection and explainable insights into decision rationale.

II. Related Work

A. COVID-19 Detection in CT images

Silva *et al.* [4] proposed an EfficientNet-based framework that addresses two key challenges in CT-based COVID-19 screening: patient-wise data leakage and dataset heterogeneity. First, instead of treating each CT slice independently, they group all slices from a patient and make a joint decision through a voting scheme—this prevents the same patient from appearing in both train and test sets. Second, they perform a cross-dataset analysis on two large public collections and show that accuracy can drop from 87.7 % to 56.2 % when models trained on one dataset are tested on another, highlighting serious generalization issues. Their work underlines the need for robust architectures and diverse data when automating COVID-19 diagnosis from CT scans.

B. Vision Transformers in Medical Imaging

Henry *et al.* [2] provide a comprehensive review of transformer models in medical imaging, tracing the evolution from natural-language transformers to pure Vision Transformers (ViT) and hybrid designs. They explain how ViTs split images into fixed-size patches and apply self-attention—originally designed for sequences—to capture long-range dependencies. While pure ViTs like the original ViT-B/16 often require large datasets, variants such as DeiT (data-efficient image Transformer) and Swin Transformer introduce inductive biases or hierarchical attention windows to improve performance on smaller medical datasets. The survey also classifies hybrids into ConvNet-like Transformers, Transformer-like ConvNets, and full Transformer-ConvNet combinations, illustrating how each balances locality and global context in tasks like classification, segmentation, and reconstruction.

C. Explainability: Grad-CAM, MoRF/AOPC

Lee *et al.* [3] analyze quantitative evaluation methods for heatmap-based explanations and introduce the Area Over the Perturbation Curve (AOPC) under the Most Relevant First (MoRF) protocol. In MoRF, the top-k pixels ranked by a heatmap (e.g., from Grad-CAM) are iteratively masked, and the drop in model confidence is recorded at each step. AOPC then computes the average confidence decrease over all masking steps, yielding a single score: the larger the AOPC, the more faithfully the heatmap highlights truly important regions. This method provides an objective, model-centric way to assess explainability, complementing visual inspection of SmoothGradCAM++ overlays.

III. Dataset & Preprocessing

A. Dataset Description

We use the SARS-COV-2 CT-Scan Dataset, publicly available on Kaggle[1]. This collection contains 1,252 CT images from COVID-positive patients and 1,230 CT images from non-COVID subjects, all provided as 2D slices. We split each class independently—70% for training (876 COVID, 860 non-COVID), 15% for validation (188 COVID, 185 non-COVID), and 15% for testing (188 COVID, 185 non-COVID)—using a custom script that shuffles and allocates images per class.

B. Image Preprocessing

Every CT slice is first resized to 224×224 pixels Resize((224,224))). For CNN inputs, we convert to single-channel grayscale, then to tensor ToTensor() and normalize with mean=[0.5], std=[0.5]. For ViT inputs, we replicate the grayscale tensor across three channels (e.g. via Lambda(x: x.repeat(3,1,1))) and normalize all channels with ImageNet statistics (mean=[0.485,0.456,0.406], std=[0.229,0.224,0.225]) to maintain consistency.

C. Data Augmentations

During training—and only then—each resized slice is flipped left–right with 50 % probability and rotated by up to ±10° to simulate anatomical and positioning variability. These simple, on-the-fly transformations help prevent overfitting and improve model robustness. Validation and test sets undergo only the fixed preprocessing pipeline described above.

IV. Models

A. Model Architectures

- CustomCNN: A lightweight network of approximately 1.2 million parameters, built from scratch for 224×224 CT slices. It consists of four Conv–BatchNorm–ReLU blocks with channel sizes [32, 64, 128, 256], each followed by max-pooling, then a global average pooling layer, a dropout at 30 %, and a final linear classifier for two outputs.
- ResNet18: The standard 18-layer residual network pretrained on ImageNet with approximately 11.2 million parameters, modified to accept one-channel input by replacing its first convolution (kernel 7×7, stride 2, pad 3). The original 1000-class head is swapped for a two-unit linear layer, with all other pretrained weights retained for transfer learning on CT data.
- ViT-B/16: The Vision Transformer "Base" model with 16×16 patches and approximately 86 million parameters, loaded with ImageNet-1K weights. We replace its 1000-class head with a two-unit linear classifier. Inputs are 224×224 grayscale slices replicated into three channels to match the pretrained patch embedding.

[1](https://www.kaggle.com/datasets/plameneduardo/sarscov2-ctscan-dataset)

B. Training Protocol

All models train using AdamW at a fixed learning rate of 1×10^{-4}, standard cross-entropy loss with label smoothing of 0.05, a batch size of 32, and four data-loading workers. We apply early stopping on validation loss with a patience of 7 epochs—if loss does not improve for 7 straight epochs, training halts and the best weights are restored.

CustomCNN is trained for up to 70 epochs, while ResNet18 and ViT-B/16 are each trained for a maximum of 20 epochs. No learning rate scheduler is used.

C. Explainability Techniques

To open the "black box" of our models, we apply SmoothGradCAM++ to the final convolutional layer of each network, producing a fine-grained heatmap that highlights the pixels most responsible for a given prediction. In practice, we normalize these activations to $[0, 1]$, resize them back to 224×224, and blend them over the original image in a Viridis colormap to produce intuitive overlays.

We then quantify the faithfulness of these explanations using the Most–Relevant–First (MoRF) protocol and the Area Over the Perturbation Curve (AOPC). Let

$$p_0 = f(\mathbf{x}) \tag{1}$$

be the model's predicted probability for the true class on the original image $\mathbf{x}$. We rank all pixels by their SmoothGradCAM++ score and, for $k = 1, \ldots, L$ (with $L = 20$), mask out the top $\frac{k}{L}$ fraction of pixels—replacing them with a mid-gray value—to obtain a perturbed image $\mathbf{x}_k$ and its corresponding probability

$$p_k = f(\mathbf{x}_k). \tag{2}$$

MoRF refers to this process of removing the Most Relevant pixels First. We then define the AOPC score as

$$\mathrm{AOPC} \;=\; \frac{1}{L}\sum_{k=1}^{L}\big(p_{k-1} - p_k\big), \tag{3}$$

which measures the average drop in confidence as high-importance pixels are removed. A larger AOPC indicates that the heatmap correctly identifies regions critical to the model's decision.

V. Results

A. Classification Performance

Table I summarizes test-set metrics for all three models.

TABLE I
TEST-SET PERFORMANCE COMPARISON

Model	Accuracy	Precision	Recall	macro F1-Score
CustomCNN	0.9786	0.9732	0.98	0.9784
ResNet18	0.9893	0.9839	0.9945	0.9892
ViT-B/16	0.9839	0.9837	0.9837	0.9837

CustomCNN delivers solid performance with 97.86 % accuracy and a macro F1-score of 0.9784, though its recall (0.9800) is slightly lower, indicating it misses a few positive COVID cases. ResNet18 leads all models, achieving 98.93 % accuracy and a macro F1-score of 0.9892 by combining very high precision (0.9839) with outstanding recall (0.9945), which minimizes both false negatives and false positives. ViT-B/16 also performs strongly (98.39 % accuracy, macro F1 = 0.9837), matching CustomCNN's precision while improving recall to 0.9837, demonstrating that transformer-based features transfer well to CT data despite limited training samples.

B. Confusion Matrices

Figure 1 shows the confusion matrices for each model on the test set.

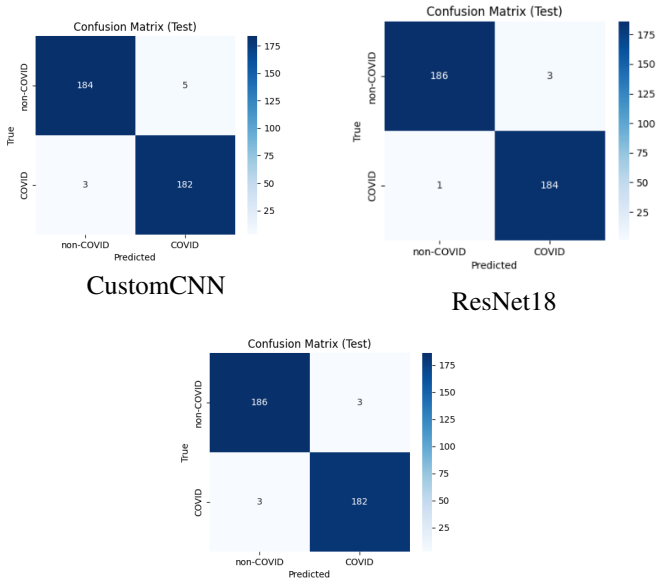


Fig. 1. Test-set confusion matrices for CustomCNN (top-left), ResNet18 (top-right), and ViT-B/16 (bottom).

CustomCNN makes 5 false positives (non-COVID misclassified as COVID) and 3 false negatives (COVID missed), indicating slightly more caution in positive predictions. ResNet18 reduces these errors to only 3 false positives and 1 false negative, achieving near-perfect separation on the test set. ViT-B/16 shows balanced behaviour with 3 false positives and 3 false negatives, suggesting similar sensitivity to CustomCNN but higher precision than CustomCNN.

C. Loss Curves

Figure 2 shows training (blue) and validation (orange) loss across epochs.

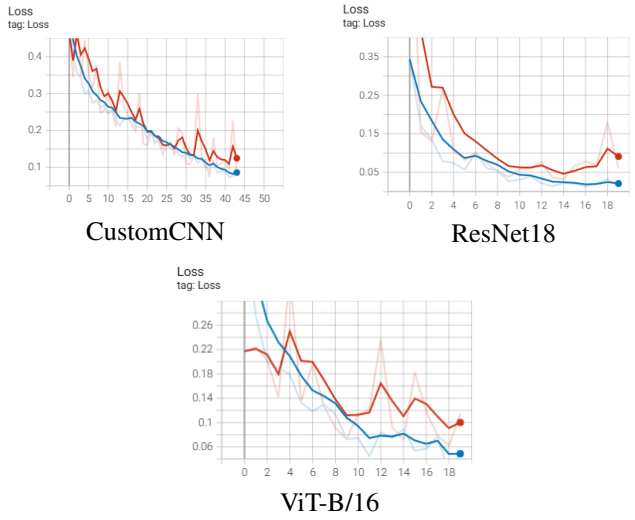
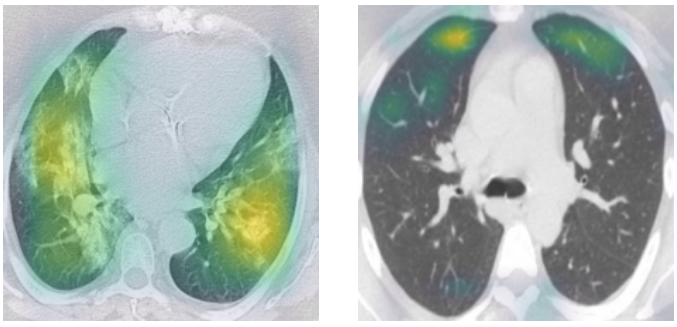


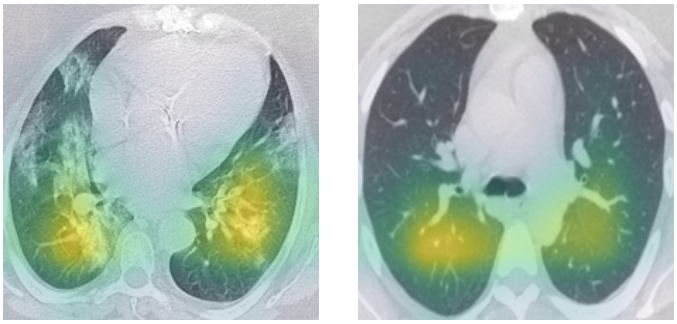
Fig. 2. Training (blue) and validation (orange) loss for each model.

All models steadily decrease both training and validation losses, indicating stable convergence. CustomCNN takes until around epoch 40 to plateau, reflecting its longer 70-epoch schedule. ResNet18 and ViT-B/16 converge faster, by epoch 18–20, thanks to pretrained weights. Minimal gaps between training and validation curves demonstrate effective regularization and low overfitting across models.

D. Grad-CAM Visualizations

Figure 3 presents SmoothGradCAM++ overlays for both a correctly classified COVID case (left) and a non-COVID case (right) across all three models.





ResNet18 (COVID) ResNet18 (non-COVID)

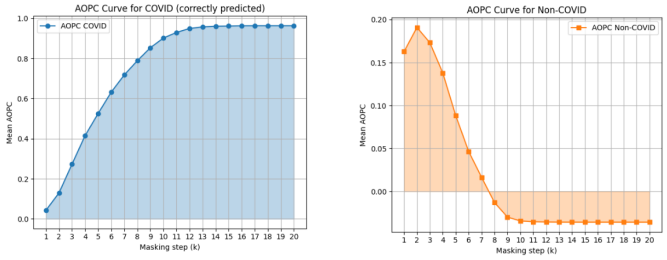
ViT-B/16 (COVID) ViT-B/16 (non-COVID)

Fig. 3. SmoothGradCAM++ heatmaps for correctly classified COVID and non-COVID examples by all the three models.

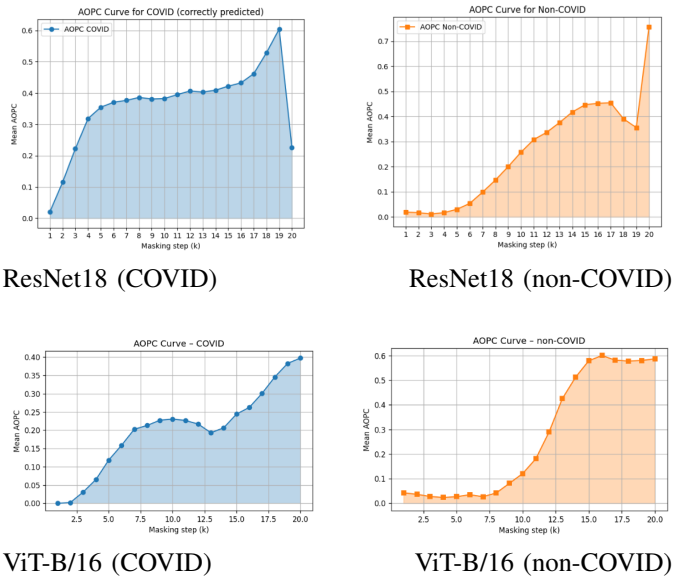
CustomCNN highlights broad peripheral ground-glass regions in the COVID example, matching radiological signs, while showing minimal activation in the non-COVID case. ResNet18 produces sharper, more concentrated hotspots exactly over the lesion areas for COVID and virtually no spurious activations for non-COVID, indicating precise localization. ViT-B/16 heatmaps are smoother and more diffuse but still focus on the correct pathological regions, demonstrating that transformer features transfer effectively to CT data despite lower spatial resolution.

E. MoRF & AOPC Analysis

We assess explanation faithfulness both visually (MoRF) and quantitatively (AOPC). Figure 4 shows AOPC curves (mean drop in predicted probability vs. masking step k), separately for COVID (left) and non-COVID (right). Figure 5 shows MoRF curves (mean class probability vs. fraction of top-importance pixels removed).



CustomCNN (COVID) CustomCNN (non-COVID)



ResNet18 (COVID) ResNet18 (non-COVID)

ViT-B/16 (COVID) ViT-B/16 (non-COVID)

Fig. 4. AOPC curves showing the average drop in true-class probability versus masking step k, for correctly classified COVID (left) and non-COVID (right) cases.

CustomCNN: The AOPC curve for COVID (top-left) climbs steeply to near 1.0 by step 10, showing that masking the most important pixels rapidly destroys the model's confidence. In contrast, the non-COVID curve (top-right) drops below zero and remains flat—CustomCNN has far fewer parameters and appears to learn only the presence of COVID features, without modeling non-COVID patterns. As a result, removing its "important" pixels actually increases or leaves unchanged its confidence on negative cases, indicating it does not learn explicit non-COVID cues.

ResNet18: Both COVID and non-COVID AOPC curves rise smoothly, peaking around steps 8–12 for COVID and steps 15–18 for non-COVID. This balanced behavior shows ResNet18 has learned distinct features for both classes: masking either positive or negative evidence consistently reduces confidence in the correct class.

ViT-B/16: The ViT's COVID AOPC (bottom-left) increases more gradually, reaching about 0.40 by step 20, while its non-COVID AOPC (bottom-right) climbs sharply after step 8 to around 0.60. This suggests the transformer spreads importance more diffusely across the image, but still captures both COVID and non-COVID evidence—albeit with slightly lower sensitivity than ResNet18.

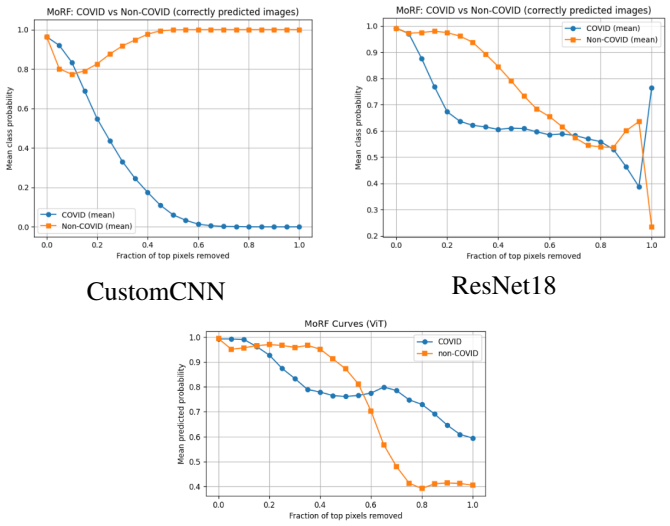


Fig. 5. MoRF curves for correctly classified COVID examples: mean predicted probability versus fraction of the top-importance pixels removed.

For CustomCNN, the COVID probability drops sharply to zero after masking about 30 % of the top pixels, while the non-COVID curve initially dips then rises back toward one—showing that this small model has learned primarily COVID features and lacks explicit negative-class cues. ResNet18 exhibits a more balanced behavior: COVID confidence declines steadily but remains above 0.5 until late masking steps, and non-COVID confidence stays high until over 50 % of pixels are removed, indicating robust, distributed feature learning for both classes. ViT-B/16 shows the slowest, smoothest decay for both curves, with COVID probability only falling below 0.5 after approximately 80 % removal, reflecting its broader attention span across the image while still faithfully capturing key regions.

VI. Discussion

Our comparison of three architectures underscores the power of transfer learning for COVID-19 detection on limited chest CT data. The CustomCNN, trained from scratch, reached 97.9 % accuracy but required nearly 40 epochs to converge and exhibited one-sided explanations—its MoRF/AOPC curves revealed it learned only "COVID-positive" patterns and ignored explicit non-COVID cues. In contrast, both ResNet18 and ViT-B/16, fine-tuned from ImageNet, converged in under 20 epochs and achieved higher accuracy (98.9 % and 98.4 %, respectively) with balanced heatmaps and perturbation curves for both classes. MoRF/AOPC analysis showed that ResNet18 and ViT-B/16 both focus on ground-glass opacities when predicting COVID and on clear lung regions when predicting non-COVID. ResNet18 produced the sharpest, most localized heatmaps, and its AOPC/MoRF curves revealed a steady, proportional loss of confidence as important pixels were removed—clear evidence of balanced feature learning. ViT-B/16 yielded smoother, more diffuse attention maps but still correctly highlighted those clinically relevant areas. In contrast, CustomCNN demonstrated one-sided behavior: its heatmaps focused on COVID patterns only, and its non-COVID AOPC curve even increased at early masking steps, indicating it lacked negative-class feature learning. Key limitations include our dataset size (1,252 COVID and 1,230 non-COVID slices), which may not capture the full spectrum of COVID presentations or scanner variations. All data come from a single public source; multi-institutional studies would improve generalization. Finally, our binary setup does not address other thoracic conditions that share similar imaging patterns, pointing to future work on multi-class classification and self-supervised pretraining on large CT repositories.

VII. Conclusion & Future Work

In this work, we compared a from-scratch CustomCNN, a fine-tuned ResNet18, and a fine-tuned ViT-B/16 for binary COVID-19 detection on chest CT slices. Our experiments on a balanced dataset of 2,482 images showed that transfer learning with ResNet18 achieves the best accuracy (98.9 %) and the most faithful explanations, producing sharp Grad-CAM heatmaps and proportional MoRF/AOPC curves for both COVID and non-COVID cases. ViT-B/16 also delivered strong performance (98.4 %) and broader but reliable attention maps, while the CustomCNN reached 97.9 % accuracy but learned only COVID features, failing to model clear-lung cues.

Next steps could include exploration of self-supervised pretraining directly on large unlabeled CT volumes, which could provide richer representations than ImageNet weights for medical images. Extension of our framework to multi-class classification—distinguishing COVID from other lung diseases such as pneumonia or fibrosis—would test the models' ability to learn diverse pathological patterns. Finally, ensembling multiple architectures that may further boost accuracy and robustness, combining the precise localization of ResNet with the global context of transformers. These directions aim to improve both diagnostic performance and explainability for real-world clinical deployment.

References

[1] F. M. Shah, S. K. S. Joy, F. Ahmed, T. Hossain, M. Humaira, A. S. Ami, S. Paul, M. A. R. K. Jim, and S. Ahmed, "A comprehensive survey of covid-19 detection using medical images," *SN Computer Science*, vol. 2, no. 6, p. 434, 2021.

[2] E. U. Henry, O. Emebob, and C. A. Omonhinmin, "Vision transformers in medical imaging: A review," *arXiv preprint arXiv:2211.10043*, 2022.

[3] J. Lee, H. Cho, Y. J. Pyun, S.-J. Kang, and H. Nam, "Heatmap assisted accuracy score evaluation method for machine-centric explainable deep neural networks," *IEEE Access*, vol. 10, pp. 64 832–64 849, 2022.

[4] P. Silva, E. Luz, G. Silva, G. Moreira, R. Silva, D. Lucio, and D. Menotti, "Covid-19 detection in ct images with deep learning: A voting-based scheme and cross-datasets analysis," *Informatics in medicine unlocked*, vol. 20, p. 100427, 2020.