Bulletin of Electrical Engineering and Informatics
Vol. 15, No. 6, December 2026, pp. 5033~5045
ISSN: 2302-9285, DOI: 10.11591/eei.v15i6.13484

Interpretable vision transformer with dual loss and Grad-CAM for iris-based early lung damage detection

Hung Ho-Dac[1,4], Van-Len Vo[2], Trong-Thua Huynh[3]
[1]Digital Transformation Center, Thu Dau Mot University, Ho Chi Minh City, Vietnam
[2]Institute of Digital Technology, Thu Dau Mot University, Ho Chi Minh City, Vietnam
[3]Faculty of Information Technology II, Posts and Telecommunications Institute of Technology, Ho Chi Minh City, Vietnam
[4]Department of Graduate Training, Posts and Telecommunications Institute of Technology, Ha Noi City, Vietnam

Article Info

Article history:

Received May 26, 2026
Revised Sep 11, 2026
Accepted Oct 4, 2026

Keywords:

Gradient-weighted class activation mapping
Iridology
Lung damage
Pre-screening
Vision transformer

ABSTRACT

Iridology infers organ health from the appearance of the iris; its clinical record is contested, but iris photography combined with learned representations allows the underlying question-whether disease-correlated structure exists in iris images-to be tested against external clinical labels. We present an interpretable framework combining a vision transformer (ViT) trained with a dual cross-entropy and triplet objective with a quantitative evaluation of its gradient-weighted class activation mapping (Grad-CAM) explanations, for binary classification of lung damage from iris images labelled by independent clinical diagnosis. On a held-out set of 356 single-participant images the model reaches 0.9249 precision and 0.9242 recall, matching a ResNet-50 baseline and a single-loss ViT to within two images at 21% lower datacentre latency. The principal contribution is the measurement of interpretability rather than its assertion: true-positive saliency maps are causally faithful, reproducible and class-discriminative, while their 0.9315 overlap with false-positive maps bounds what they may be taken to mean-they show where the model finds positive evidence, not where pathology is. We additionally audit acquisition confounding and find that image-quality statistics alone separate the classes at area under the receiver operating characteristic curve (AUROC), establishing a floor that iris-based screening studies should report and that, to our knowledge.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Hung Ho-Dac
Digital Transformation Center, Thu Dau Mot University
06 Tran Van On street, Phu Loi ward, Ho Chi Minh City, Vietnam
Email: hunghd@tdmu.edu.vn

1. INTRODUCTION

Iridology, the practice of inferring organ health from the iris, has a uniformly negative record in controlled clinical evaluation [1]–[3]. What has changed is not the evidence for the practice but the instrument: iris photography paired with learned representations allows the underlying question-whether any systematic, disease-correlated structure exists in iris images-to be tested against external clinical labels, with the failure modes measured rather than assumed [4], [5]. The question matters because post-coronavirus disease 2019 (post-COVID-19) pulmonary sequelae are prevalent and under-detected, and those least likely to be detected have least access to spirometry, which needs a forced expiratory maneuver unreliable in the frail and elderly, and to radiography, which involves radiation and cost an unselected referral does not justify. An inexpensive, non-radiative photograph is therefore attractive as a triage layer upstream of both.

Machine learning on iris images has been applied to pulmonary disease-Gabor descriptors with a support vector machine (SVM) [6], a voting ensemble [7], mobile-scale convolutional neural networks (CNNs) [8], a residual network with recurrent attention [9], and to systemic conditions such as diabetes [10]. Özbilgin and Kurnaz [11] form a methodologically distinct line, correlating the sodium-ring grey level with serum cholesterol at about 89% agreement, reaching 93% accuracy for coronary artery disease [12], and combining iris features with bio-demographic profiles in a stacking ensemble [13]. These bear on the present work in three ways: clinically relevant information can be extracted from standardized iris photographs; analysis is delimited to a region rather than the whole iris; and objective clinical measures serve as the reference standard rather than iridological assessment.

Vision transformers (ViTs) [14] model relationships between image patches with self-attention, capturing long-range dependencies that convolutional networks express less efficiently-an advantage where diagnostic structure is distributed rather than local. In lung imaging, Matsuyama *et al*. [15] compared a ViT against ResNet-50 on computed tomography (CT) and Malaviya *et al*. [16] proposed lightweight vision transformer (LViT); transformers have also been applied to iris biometrics [17], [18]. Applying such a model to patients demands an account of what the network responded to. The dominant post-hoc technique is gradient-weighted class activation mapping (Grad-CAM) [19]; alternatives include class-aware layer relevance for transformers [20], class-agnostic attention rollout [21], and the perturbation-based local interpretable model-agnostic explanations (LIME) [22] and Shapley additive explanations (SHAP) [23]. Subsubsection 2.6.1 sets out the choice among them.

Three deficiencies run through this literature and they compound. First, interpretability is asserted rather than measured: where saliency is shown, including in our own prior work [9], the evidence is that the highlighted region looks plausible, and plausibility cannot separate a map marking the evidence the model used from one merely coinciding with expected anatomy; no study reports a faithfulness, stability or class-discriminativeness measure. Second, no study reports the variance of its accuracy-every result in Table 1 is a point estimate from one partition, several on cohorts of a few hundred subjects. Third, no study audits whether its accuracy is attributable to the iris at all: if the two classes were captured in different sessions, by different operators or on different equipment, a systematic quality difference becomes a shortcut no saliency method can expose, because it spans the whole patch. A fourth, architectural gap follows: ViTs have not been applied to iris-based pulmonary pre-screening, where the patch-token grid supplies the spatially resolved representation a quantitative saliency evaluation requires.

Table 1. Comparison with prior iris-based lung screening studies

Study	Target	Data size	Method	Reported performance	Interpretability
Bansal *et al.* [6]	Obstructive lung disease	100 subjects	Gabor filters+SVM	88.0% accuracy	None
Rehman *et al*. [7]	Obstructive lung disease	529 subjects	Ensemble ML and voting	97.6% accuracy	None
Huynh *et al.* [9]	Lung damage	10,866 images	ResNet+BiLSTM, self-attention	96.29% accuracy, AUC 0.9923	Grad-CAM, qualitative
This study	Lung damage	3,488 images	ViT with CEL+TL	0.9249 precision, 0.9242 recall	Grad-CAM, quantified; confound audited

This study addresses all six: i) saliency is assessed quantitatively: deletion and insertion against a random order test whether the highlighted regions are causal, Gini and top-decile mass measure focalization, and inter-group overlap against a split-half reliability ceiling tests class-discriminativeness; the protocol is released as an attributor-agnostic package, ii) we audit acquisition bias by asking how far image-quality statistics alone separate the classes, and report that area under the receiver operating characteristic curve (AUROC) beside the model's, iii) we compare three configurations under one pipeline with cross-validation, confidence intervals, paired effect sizes and equivalence testing, reporting parity under a stable schedule together with an optimization failure in one fold, rather than claiming equivalence at a power the design does not support, iv) we test the mechanism of the dual loss and identify the separation ratio rather than F1 as the criterion for selecting the triplet margin, v) labels are independent clinical diagnoses, and vi) we treat the chart-derived sector as an assumption under test and specify the experiment that would falsify it (subsubsection 2.3.1).

2. METHOD

2.1. Dataset and clinical ground truth

Analysis used 3,488 iris images, one per participant, from a collaborating clinical centre; the region of interest (ROI) patch is the unit of analysis. The classes are close to balanced, 1,747 positive and 1,741 negatives.

Class membership was assigned by independent clinical diagnosis, made by qualified practitioners using standard respiratory investigation and recorded before any iris photograph was taken; iridology played no part in it. The protocol was approved by the institutional ethics committee (Thu Dau Mot University-TDMU), all participants gave written informed consent, and images were de-identified at capture. A data protection agreement prevents disclosure of the institution, the acquisition period, the instrument and participant demographics, and prevents release of the raw images; what we can disclose is the class counts above, the partition sizes, the quality statistics of subsection 3.4 and all analysis code. Subsection 3.8 treats the consequences as a limitation rather than a solved problem.

2.2. Image acquisition and preprocessing

Images were captured with an iridoscope under one standardized procedure: seated participants, head stabilized on a chin rest, controlled lighting, trained operators following written guidance; the instrument falls under the agreement above. Images were excluded for motion blur, specular reflection on the lung sector, substantial eyelid or eyelash occlusion, or off-axis gaze, and the 3,488 patches analyzed are those that passed screening. Preprocessing comprised reflection removal, pupil and limbus localization, a polar-sector crop, resizing to 224×224 with bicubic interpolation, and ImageNet normalization. Training augmentation was random horizontal flip (p=0.5), slight rotation and color jitter; none was applied to validation or test.

2.3. Region of interest extraction

Iridology charts widely reproduced in the practitioner literature [24], [25] place the lung zone between the pupillary and ciliary edges, from 2h to 4h in the right iris and 8h to 10h in the left (Figure 1). The pupil and outer iris boundary were located by circular Hough transform and a polar-sector mask applied to extract the ROI patches.

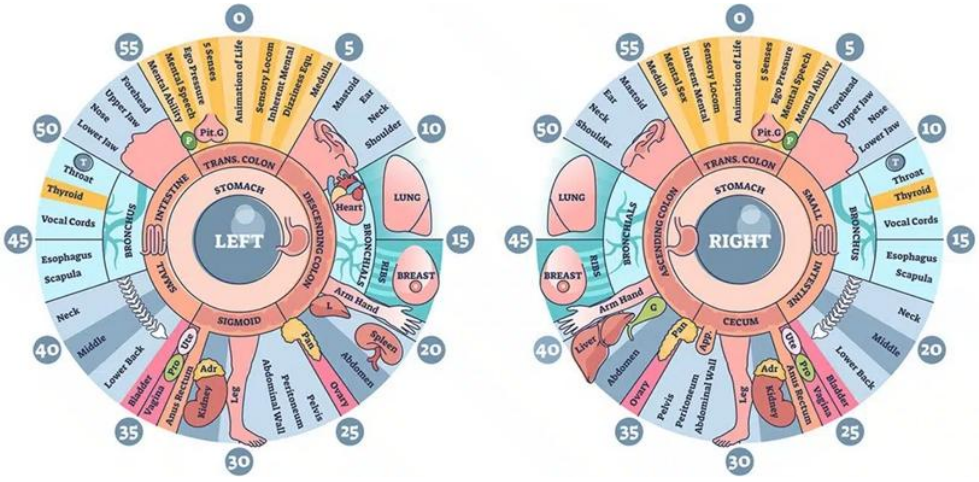


Figure 1. The chart-derived lung sector: the shaded 2h-4h (right iris) and 8h-10h (left iris) sectors

Figure 2 presents representative lung-region ROI patches and the distributions of image-quality statistics used to characterize the extracted patches. Figure 2(a) shows three representative patches, each cropped exclusively from the lung sector at a scale that preserves the visible stromal fibres and lacunar structure. Figure 2(b) presents the class-conditional distributions of three image-quality statistics across all 3,488 patches. The sharpness histograms are visibly offset between the negative and positive classes, whereas the brightness and contrast distributions overlap closely. These observations motivate the quantitative audit presented in subsection 3.4. All subsequent modelling uses ROI patches only.

2.3.1. The iridology map as a spatial prior

Restricting analysis to these sectors imports a claim from classical iridology, and that claim is an assumption rather than settled anatomy. The clinical evidence for iridology as a diagnostic practice is negative: practitioners could not distinguish renal impairment from controls [1], gallbladder detection was no better than chance [2], and a recent review [3] finds no established mechanism linking iris zones to organ pathology and notes that the principal charts assign different organs to the same sector. That review raises the objection bearing on this study: work that pre-specifies a zone from a chart and trains within it risks circular reasoning.

Circularity, however, arises when the labels are themselves chart-derived, since accuracy then measures agreement with iridology rather than with disease. That is not the design here: class membership was fixed by independent clinical diagnosis recorded before photography, so the chart contributes only a spatial prior-which pixels the network is shown-while the target is external to iridology. The restriction is therefore a falsifiable hypothesis: were the sector uninformative, F1 would sit near the 0.5 chance level of this balanced cohort. It does not; it is 0.9242 on held-out data.

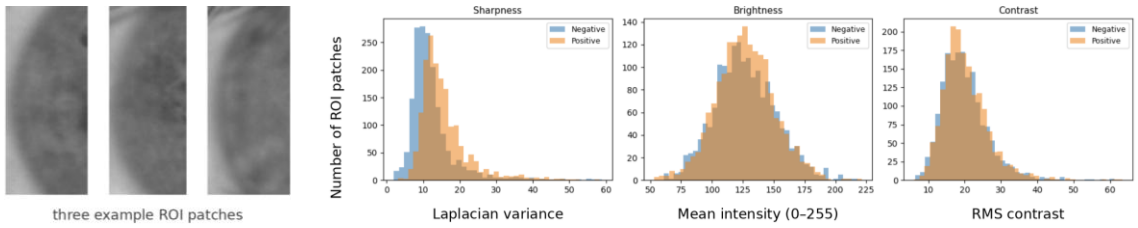


Figure 2. Representative lung-region ROI patches and class-conditional distributions of image-quality statistics: (a) three example ROI patches and (b) distributions of sharpness, brightness, and contrast for the negative and positive classes

What this does not establish is sector specificity. Three possibilities remain open: a systemic correlate of pulmonary disease may be expressed across the entire iris, in which case any sector would serve; the acquisition artefact of subsection 3.4 is a whole-image property that cropping does not remove; and the charts disagree about which sector is the lung sector. The decisive test is a control-sector ablation-retraining on an area-matched sector to which no chart assigns the lungs, and on the full annulus-which we could not run because only ROI patches were retained (subsection 3.8). It is specified in section 4 as the condition under which the iridological reading should be withdrawn. Until then our finding is narrower than this literature usually claims: a clinically diagnosable pulmonary signal is recoverable from a fixed annular sector under this acquisition protocol. That is a claim about iris imaging, not a validation of iridology.

2.4. Vision transformer architecture

The backbone is a standard ViT [14], the final class-token representation feeding a linear head. We use ViT-Base/16 initialized from ImageNet-1k weights and fine-tuned end-to-end (Table 2). The 16×16 patch size yields a 14×14 token grid at 224×224 input; a larger patch would flatten the radial and lacunar textures that carry the signal, a smaller one would grow attention cost quadratically without benefit at this dataset size. All other parameters were left at standard values and not tuned, so that the effect of the training objective alone could be observed.

Table 2. Architecture and training configuration for the three compared models

Setting	ResNet-50 baseline	ViT (CEL)	ViT (CEL+TL)
Backbone/initialization	ResNet-50, ImageNet-1k	ViT-Base/16, ImageNet-1k	ViT-Base/16, ImageNet-1k
Input/patch size/tokens	224×224/—/—	224×224/16×16/14×14	224×224/16×16/14×14
Layers/heads/dim./MLP ratio	50 conv. layers	12/12/768/4	12/12/768/4
Classification/embedding head	Linear (→ 2)/—	Linear (→2)/—	Linear (→2)/Linear (768 → 256)+L2
Optimizer/schedule/LR	Adam/cosine/1×10^{-4}	Adam/cosine/1×10^{-4}	Adam/cosine/1×10^{-4}
Batch/epochs/weight decay	32/30/1×10^{-4}	32/30/1×10^{-4}	32/30/1×10^{-4}
Triplet margin α/weight λ	—	—	0.5/0.3
Early stopping/seed	Best val. F1/42	Best val. F1/42	Best val. F1/42

2.5. Dual loss strategy

Training employs a dual loss. Cross-entropy loss (CEL) optimizes the decision boundary but does not structure the embedding space, so a triplet loss (TL) is applied concurrently to L2-normalised embeddings from a projection head. For an anchor a, a same-class positive p and an opposite-class negative n:

$$LTL = max(\|f(a) - f(p)\|_2^2 - \|f(a) - f(n)\|_2^2 + \alpha, 0) \quad (1)$$

where α is the margin, set to 0.5. The total objective is the weighted sum:

$$L = LCEL + \lambda \cdot LTL \quad (2)$$

with lambda=0.3 weighting the terms. Triplets are mined batch-hard within each stratified mini-batch of 32, because the informative triplets here are opposite-class patches that look coarsely similar, whereas random sampling yields mostly trivially satisfied triplets carrying no gradient; the triplet term is switched on after two warm-up epochs. The classification head reads the pre-normalization features rather than the projected embedding, a detail that determines the result of subsection 3.2. The margin alpha=0.5 and weight lambda=0.3 were fixed a priori from common practice and not tuned on this dataset; no other configuration was trained, so the held-out figures carry no selection effect, and their consequences are examined post hoc in subsection 3.2.

2.6. Gradient-weighted class activation mapping interpretability and its quantitative evaluation

We apply Grad-CAM [19] to the proposed model. The final self-attention layer produces patch tokens arranged in a spatial grid, treated as analogous to CNN feature maps: the gradient of the target-class logit is backpropagated to that layer's output, per-patch gradient magnitudes serve as importance weights, and the map is bilinearly up sampled to ROI resolution.

2.6.1. Choice of attribution method

Grad-CAM was chosen against three alternatives on grounds that follow from what the maps must do here. Attention rollout [21] is native to transformers but class-agnostic, returning one map per image irrespective of target class; since the principal analysis compares maps between prediction outcomes, it would make every such comparison degenerate. Class-aware transformer attribution does exist-the layer-relevance formulation of Chefer *et al*. [20] and section 4 identifies it as the primary future comparison, deferred because it requires re-instrumenting every attention block. LIME [22] and KernelSHAP [23] assume a meaningful partition into interpretable units, supplied on natural images by super pixels aligned to object boundaries; a polar-sector patch has none, so any partition is arbitrary, and both need 10^3-10^4 forward passes per image against one forward and one backward pass. Applying Grad-CAM to a transformer is nonetheless a transplant, since layer normalization and class-token routing give no guarantee that gradient magnitudes over patch tokens behave as over a convolutional map-which is why the maps are evaluated rather than asserted. The protocol is attributor-agnostic, so any alternative can be scored on the same axes with the released code.

An intersection over union (IoU) against a pre-specified anatomical region is not meaningful here, because the patches are already cropped to the lung sector, so the reference region is the whole patch and the overlap would depend only on the binarization threshold. We therefore assess three axes. Deletion and insertion remove or reveal pixels in decreasing saliency order and record the class probability at each step: a faithful map gives a low deletion area under the curve (AUC) and a high insertion AUC, with a random pixel order as control [26]. Gini and top-decile mass measure concentration. Inter-group IoU between mean maps, referenced to a split-half reliability ceiling computed within the true-positive group, tests whether attention differs systematically between prediction outcomes.

Saliency cannot detect a shortcut spanning the whole patch, so we test for one independently. For each image we compute three statistics carrying no iris-texture information-Laplacian variance, mean intensity, root mean square (RMS) contrast-compare their class-conditional distributions with Cohen's d, then fit a logistic regression on those three alone under the protocol applied to the image models. The gap between the two AUROC values bounds how much of the performance is not attributable to acquisition.

2.7. Experimental setup

Because each image belongs to a different participant, participant-level and image-level splitting are equivalent and no participant can appear in more than one partition, which removes the subject-level leakage that would otherwise make our estimates overoptimistic. Splitting was stratified on the class label with a fixed seed of 42 (Table 3). We additionally ran stratified five-fold cross-validation over the combined training and validation sets (3,132 images), reporting per-fold metrics as mean +/- standard deviation with a 95% t-interval on four degrees of freedom; five folds characterize variance adequately but give limited power for comparing configurations.

Table 3. Data partition, by image and by participant

Subset	Proportion (%)	Images	Positive	Negative	Individuals
Training	75	2,612	1,312	1,300	2,612
Validation	15	520	260	260	520
Test	10	356	175	181	356
Total	100	3,488	1,747	1,741	3,488

Three configurations were compared: a ResNet-50 baseline under cross-entropy; the ViT of subsection 2.4 under cross-entropy; and the proposed ViT under the combined objective. All share the split, augmentation, batch size and epoch budget, so differences are attributable to architecture and loss (Table 2). Experiments ran on one NVIDIA A100-SXM4 (80 GB) with PyTorch 2.11.0 (CUDA 12.8) and timm 1.0.28, CUDA deep neural network library (cuDNN) deterministic, seed 42. Evaluation prioritizes recall alongside precision and F1, because a missed positive is the more serious error in screening. The full workflow is released as executable code with a pinned environment and a synthetic ROI dataset on which the pipeline runs end to end (see data availability).

3. RESULTS AND DISCUSSION

3.1. Classification performance

Table 4 compares the three configurations on the held-out test set. The proposed ViT (CEL+TL) reaches F1 0.9242, indistinguishable from the single-loss ViT (0.9213) and the ResNet-50 baseline (0.9269); the spread is 0.0056 F1, two images at this test-set size. The diagnostic information is therefore present in the sector the chart associates with the lungs and is extractable by conventional architectures at a level comparable to the best reported for iris-based lung screening, our 0.9242 exceeding the 0.9026 of Ho-Dac *et al*. [8] on a larger dataset, though the two are not directly comparable. The dual-loss advantage is +0.0029 F1, one image, and the cross-validation ranks the two the other way, so we treat all three as performing at the same level.

Table 4. Experimental results of the models on the held-out test set (n=356)

Model	Precision	Recall	F1-Score	Loss
ResNet-50 baseline	0.9290	0.9270	0.9269	0.16
ViT (CEL)	0.9223	0.9213	0.9213	0.18
ViT (CEL+TL)	0.9249	0.9242	0.9242	0.17

3.2. Statistical validation and embedding geometry

Point estimates from one partition cannot establish that a difference is robust, so all three configurations were also cross-validated (Table 5), with paired tests of fold-wise F1 in Table 6. We do not interpret the non-significant p-values as evidence of equivalence: with five folds a paired t-test has minimal power, and the 95% interval against ResNet-50 spans [-0.2237, 0.0932], wide enough to contain both a large advantage and a large deficit. The comparison is inconclusive, and the effect sizes (Cohen's d=-0.485 and -0.512) point, if anything, against the proposed configuration. Upper interval bounds in Table 5 are truncated at the attainable maximum F1=1.000; the untruncated values are 1.0016 and 1.0408, an artefact of a t-interval on a bounded metric with four degrees of freedom.

Table 5. Stratified five-fold cross-validation: mean +/- standard deviation and 95% confidence interval

Model	Precision	Recall	F1-Score	95% CI of F1
ResNet-50 baseline	0.9473±0.0048	0.9470±0.0048	0.9470±0.0048	[0.9410, 0.9529]
ViT (CEL)	0.9088±0.0737	0.9071±0.0760	0.9070±0.0762	[0.8124, 1.0000]
ViT (CEL+TL)	0.8933±0.1033	0.8844±0.1221	0.8817±0.1281	[0.7226, 1.0000]

Table 6. Paired comparison against the proposed model on fold-wise F1 (five folds)

Comparison	Mean ΔF1	95% CI	Paired t (p)	Wilcoxon (p)	Cohen's d
ViT (CEL+TL) vs. ViT (CEL)	−0.0253	[−0.0901, 0.0395]	0.3395	0.3125	−0.485
ViT (CEL+TL) vs. ResNet-50	−0.0653	[−0.2237, 0.0932]	0.3164	0.1250	−0.512

The variance is confined to one fold, where the single-loss ViT reaches F1 0.7711 and the dual-loss ViT 0.6527 while ResNet-50 is unaffected at 0.9457; across the other four folds the three score 0.9473, 0.9410 and 0.9389 with standard deviations below 0.007. The distribution is bimodal with no intermediate values and only the transformers are affected-the signature of an optimization failure, not a representational deficit. A single learning rate of 1×10^{-4} was used for all three to keep the comparison controlled, which suits a 24.4 M-parameter CNN but is aggressive for an 86.2 M-parameter transformer, for which layer-wise learning-rate decay is standard. We retain the uncorrected result as primary, since re-running one fold under a different schedule would break the controlled comparison.

We therefore add a secondary analysis excluding that fold, flagged as post hoc since it was identified after the results were seen. Over the four stable folds the mean difference against the single-loss ViT is -0.0021

F1 and against ResNet-50 -0.0084 F1, twelve-fold and eight-fold smaller than over the full set, the latter three images at the held-out test size. We also report a two one-sided tests (TOST) procedure [27] against a pre-specified equivalence margin of +/- 0.02 F1, about seven images at n=356. The claim is bounded accordingly: on four of five folds and on the held-out partition a ViT fine-tuned on 2,612 images matches a convolutional baseline to within a few images, while on one fold it fails to optimize under a shared learning rate-transformer fine-tuning at this scale is schedule-sensitive in a way the baseline is not. The evidence supports parity under a stable schedule, not superiority and not equivalence at adequate power.

Subsection 2.5 argued that the triplet term should compact the class manifolds while keeping them apart. Table 7 tests that claim. Compaction is large, mean intra-class distance falling from 0.5510 to 0.0046, but at the same margin the inter-class distance falls in almost the same proportion, from 1.0708 to 0.0075, so the separation ratio declines from 1.9434 to 1.6174.

The maximum distance between L2-normalised embeddings is 2.0, so alpha=0.5 demands a quarter of the representable range-satisfiable either by separating the manifolds, which is the intent, or by contracting all embeddings uniformly, which satisfies the triplet inequality equally well and is the easier path. Because the head reads pre-normalization features, that contraction is largely invisible to the classifier: the geometric component is not absent but functioning suboptimal at this margin. This also explains why a sensitivity sweep over alpha and lambda scored on F1 would have been uninformative, since a grid evaluated on a metric that moves by one image between configurations returns a flat surface within fold-to-fold noise. The informative criterion is the separation ratio. Subsequent work should sweep alpha in {0.05, 0.1, 0.2, 0.3, 0.5} against lambda in {0.1, 0.3, 0.5, 1.0}, selecting on the validation separation ratio with F1 as a constraint; a narrower margin, a soft margin or a dispersion loss are the usual remedies, all detectable by that measure.

Table 7. Embedding space geometry on the test set

Model	Mean intra-class distance	Mean inter-class distance	Separation ratio
ViT (CEL)	0.5510	1.0708	1.9434
ViT (CEL+TL)	0.0046	0.0075	1.6174

3.3. Quantitative evaluation of saliency

Grad-CAM maps for all test images were obtained from the dual-loss model with the target class set to the predicted class; they are shown in Figure 3 and measured in Table 8.

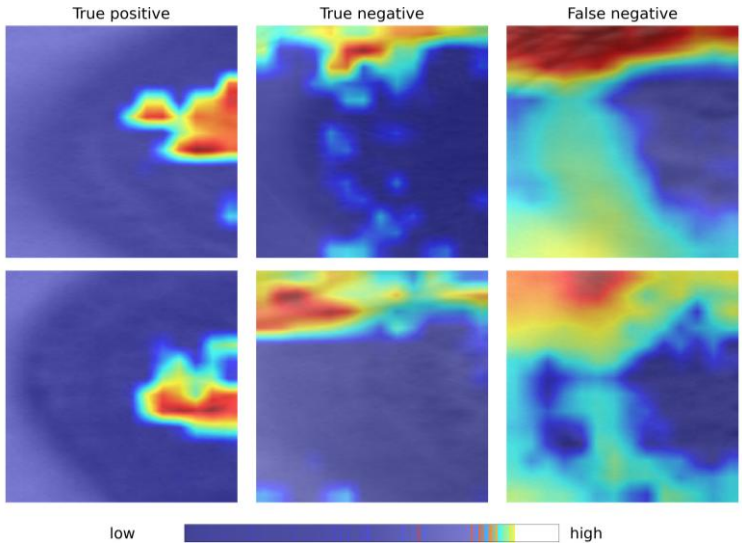


Figure 3. Grad-CAM saliency maps overlaid on ROI patches

Table 8. Quantitative saliency evaluation by prediction outcome

Prediction outcome	n	Deletion AUC	Insertion AUC	Gini	Top-decile mass
True positive	165	0.3555±0.2581	0.7349±0.1547	0.2687±0.3527	0.1793±0.2721
True negative	162	0.8522±0.1912	0.9938±0.0085	0.1865±0.0936	0.1474±0.0301
False negative	10	0.5384±0.2029	0.9754±0.0215	0.3422±0.1414	0.2030±0.0671
False positive	19	0.0224±0.0207	0.8109±0.0737	0.7114±0.0838	0.4384±0.1521
Random-order	356	0.6934±0.3180	0.7104±0.3035	-	-

For true positives the deletion AUC is 0.3555 against a random-order baseline of 0.6934: deletion in decreasing saliency order removes the evidence 0.3379 AUC faster than random deletion, a standardized difference of about 1.17. The highlighted regions therefore contain the evidence rather than merely coinciding with it-the property that cannot be confirmed by inspection and has not previously been demonstrated in this domain. True-positive maps are also denser than true-negative maps (Gini 0.2687 against 0.1865; top-decile density 0.1793 against 0.1474). The dispersion is nonetheless considerable: the Gini standard deviation for true positives, 0.3527, exceeds its mean, so a number of true-positive maps are in fact diffuse.

Binarizing each group's mean map at its 80th percentile, the split-half reliability ceiling within the true-positive group is IoU 0.9465, so the model attends to a stable region rather than image-specific artefacts. Against that ceiling, true-positive versus true-negative and versus false-negative overlaps are both 0.0000: positive and negative evidence come from different areas of the sector, and in misses the displacement is complete.

The true-positive versus false-positive comparison, at 0.9315, sits near the ceiling: the model applies the same evidential criterion whenever it predicts positive, whether or not the case is positive. Read constructively this is behavioral consistency. Read strictly it bounds the interpretability claim-the map shows where the model sees positive evidence, not whether pathology is present there-and the two can be separated only by the disaggregated outcome analysis used here.

Two rows of Table 8 expose a property of the protocol rather than the model. The random-order control is pooled over all 356 images rather than computed within each outcome group; group-specific controls would be preferable and are recommended for replication. That pooling produces the true-negative row, where deletion AUC 0.8522 sits above the random control: deletion and insertion are defined against a blurred baseline [26], and here a blurred patch is itself strong evidence for the negative class, since loss of fine structure distinguishes the classes. The reference is therefore not class-neutral, so these metrics are interpretable for the positive class and largely uninterpretable for the negative one, reported for completeness rather than as evidence; a class-neutral reference such as a per-class mean patch would be the correction. The false-positive group shows the opposite extreme, but at n=19 and n=10 for false negatives, both are descriptive only.

3.4. Acquisition-confound audit

Where a model separates the classes, it is reading either the physiology or the conditions of acquisition, and a quality shortcut spans the whole patch, so no saliency analysis can detect it. The audit identifies one artefact and clears two. Sharpness differs between classes, 15.73+/-6.72 positive against 12.32+/-6.35 negative, Cohen's d=0.522. Brightness (d=-0.006) and contrast (d=0.041) are indistinguishable, so illumination and exposure were well controlled.

Both rows of Table 9 were produced under one protocol; because it deviates slightly from subsection 2.7, the model row differs from Table 4 (F1 0.9217 against 0.9242) and the two must not be compared across. An AUROC of 0.7482 is attainable without examining the iris, so a substantial part of the class separation here can be inferred from acquisition alone, and the sharpness difference most likely reflects session or operator rather than diseased irises. The model's 0.9812 nonetheless stands well above that floor: the quality baseline mis-orders about one positive-negative pair in four, the model fewer than one in fifty. We report the artefact because 0.7482 is a floor below which no value on this dataset can be taken as evidence of an iridological signal, and any future collection must randomize or match acquisition conditions to remove it rather than merely exceed it. Prior work provides no such audit; it costs nothing and bounds the central claim of any such study.

Table 9. Class separability from image-quality statistics alone, against the proposed model

Predictor set	Precision	Recall	F1	AUROC
Image quality only (sharpness, brightness, and contrast)	0.6726	0.6660	0.6628	0.7482
ViT (CEL+TL), full ROI patch	0.9223	0.9218	0.9217	0.9812

3.5. Computational efficiency

Table 10 reports parameter count, training time, single-image latency, and peak memory, measured on the A100 of subsection 2.7 with cuDNN deterministic; CPU-only and mobile-GPU latency were not measured. The TL costs 0.2 min over the whole budget and nothing at inference, since the embedding head is not required after deployment. Against the baseline the proposed model trains in 9.5 min and infers in 5.43 ms against 14.5 min and 6.82 ms, a 21% latency reduction at the same 1.64 GB peak memory. The transformer is faster despite 3.5 times the parameters because dense attention saturates the A100's tensor cores while ResNet-50's sequential stages leave the device underutilized at this batch size. This is a property of the hardware: on a CPU or integrated mobile GPU the parameter count dominates and the ordering should reverse, so the figure is a datacentre measurement relevant to the centralized topology of subsection 3.7, not an efficiency claim for edge devices.

Table 10. Computational cost of each configuration

Model	Parameters (M)	Training time (min)	Inference latency (ms)	Peak memory (GB)
ResNet-50 baseline	24.4	14.5	6.82	1.64
ViT (CEL)	86.2	9.3	5.57	1.64
ViT (CEL+TL)	86.2	9.5	5.43	1.64

3.6. Comparison with prior iridology-based lung screening

Table 1 places these findings against previous iris-based lung screening studies. The comparison is indicative rather than controlled, since the studies differ in target condition, cohort size, reference standard, and assessment protocol. Two prior works report higher accuracy, but Bansal *et al*. [6], Rehman *et al*. [7] use 100 and 529 subjects, so one favorable partition can move the headline figure substantially, and all four report a single point estimate with no variance.

The distinctive element of this study is therefore the evidentiary bar it raises rather than the figure it reports. Prior work in this vein, including our own [9], validated Grad-CAM by visual confirmation, which does not distinguish appropriate anatomical attention from coincidentally correct attention; none reports cross-validated variance; and none audits for acquisition confounding. Since image-quality statistics alone reach AUROC 0.7482 here, a study reporting a single accuracy figure without such an audit cannot say what portion of it is due to physiology.

These findings sit within a broader reassessment of post-hoc explanation in medical imaging. Attribution maps can be insensitive to model parameters and to labels [28], and in a systematic radiological evaluation localized abnormalities less reliably than human annotation, with the discrepancy invisible to a reader judging by eye [29]. The field has divided between replacing post-hoc explanation with intrinsically interpretable models in high-stakes settings [30] and retaining it where faithfulness is measured. This study takes the second position: every interpretability claim here is attached to a measurement, and the claim the evidence does not support-that the overlay localizes pathology rather than model evidence-is stated as a bound. Disaggregating saliency statistics by confusion-matrix cell costs nothing, and a pooled evaluation would have reported the same aggregate faithfulness while concealing that true-positive and false-positive maps are nearly indistinguishable.

3.7. Clinical implications and deployment considerations

The role this framework could occupy is narrow: not a diagnostic test, and no substitute for spirometry or radiography, but a triage layer at first contact. A community health post with a hand-held iridoscope and a network connection could photograph a presenting adult in under a minute, without a trained operator or consumables, and return a stratification determining whether that person joins the spirometry queue now or at routine interval. The value lies not in accuracy relative to spirometry but in the composition of that queue: a test with high sensitivity and modest specificity raises the pre-test probability of the population reaching the confirmatory test. This is why the evaluation prioritizes recall, and why a deployed operating point should favor sensitivity rather than maximize F1.

Three properties make the modality complementary rather than competitive: no radiation, so it can be repeated at every contact; no patient effort, unlike spirometry; and single-image capture, so analysis can be centralized, which given the latency of subsection 3.5 means one server can serve many peripheral sites. Both cohorts here are close to balanced, however, which reflects the recruitment design and not any screening population: at realistic prevalence the positive predictive value would be substantially lower than the precision in Table 4. Precision and recall as reported characterize discriminative capacity, not performance in a screening programmed.

Two findings impose operational conditions on deployment. Because a measurable component of class separation here is attributable to acquisition, a deployed model will respond partly to how an image was captured, and the resulting errors would cluster by site, operator and time of day-the structure most likely to escape aggregate monitoring and to fall unevenly across populations; standardized acquisition with periodic per-site calibration is therefore a prerequisite. And because the maps identify where the model finds positive evidence rather than where pathology lies, any overlay shown to a clinician must be labelled as a statement about the model. Beyond these, multi-centre studies are needed to verify robustness to different cameras, lighting and populations, and prospective validation to estimate predictive values at realistic prevalence.

3.8. Limitations

The 3,488 images come from a single site under one acquisition configuration, with no external cohort. Every image is a distinct participant, which removes subject-level leakage but does not address population representativeness, and each item withheld under the data protection agreement has a cost. Without age and sex distributions we cannot judge whether the cohort resembles a screening population, nor whether an age-

correlated iris change such as arcus or stromal density co-varies with pulmonary status here-an unquantified confound, not a stylistic omission. Without the acquisition period and operator assignment the sharpness difference can be bounded but not attributed; without smoking history or comorbidity the positive class cannot be characterized beyond the binary label; without the instrument model the optical pipeline cannot be reproduced. Independent verification of representativeness is therefore unavailable.

Image-quality statistics alone reach AUROC 0.7482 against 0.9812 for the full model, so most of the discriminative signal is not attributable to acquisition, but the design cannot partition the two exactly and the reported performance should be read as an upper bound on what iris content contributes. Only ROI patches were retained, so quality can be characterized only after cropping and extraction cannot be re-run with alternative sector definitions.

The RoI was defined from charts whose empirical basis is contested (subsubsection 2.3.1). Our labels are independent clinical diagnoses, so the circularity affecting chart-derived-label studies does not apply, but we did not evaluate a control sector, so the lung sector is shown to be informative and not specifically informative. Separately, five folds are underpowered for comparing configurations: the non-significant differences are not evidence of equivalence, and the four-fold analysis is post hoc.

Three bounds apply to the interpretability results. Grad-CAM on a patch-token grid is an adaptation of a method designed for convolutional maps, validated here rather than assumed, but not compared against class-aware transformer attribution [20]. The deletion-insertion reference image is not class-neutral here, which makes those metrics uninterpretable for the negative class. The framework was not evaluated with clinicians, so nothing is known about whether these overlays help or mislead a human reader-the question that ultimately determines whether the term interpretable is warranted.

4. CONCLUSION

On 3,488 single-participant ROI patches a ViT-B/16 trained with a combined cross-entropy and triplet objective reached 0.9249 precision and 0.9242 recall on a held-out test set, matching a ResNet-50 baseline (0.9290/0.9270) and a single-loss ViT (0.9223/0.9213) to within two images while reducing inference time by 21% on datacentre hardware. Five-fold cross-validation was inconclusive at the available power and identified a transformer optimization failure in one fold; excluding it as a disclosed post-hoc analysis, the difference from the baseline is 0.0084 F1. We claim parity under a stable schedule, not superiority and not demonstrated equivalence. The embedding analysis shows the class manifolds contracting by two orders of magnitude, but uniformly rather than class-selectively, leaving margin selection as a diagnosable avenue for improvement.

The interpretability analysis is the principal contribution, and it is the measurement of interpretability rather than its assertion. Unlike other studies in this domain, whose validation was purely visual, ours is quantitative: true-positive maps are causally faithful (deletion AUC 0.3555 against 0.6934), reproducible (split-half IoU 0.9465) and class-discriminative (zero overlap with true-negative and false-negative maps). The same analysis supplies the bound-true-positive and false-positive maps overlap at 0.9315, so the overlay shows where the model finds positive evidence and not where pathology is-which governs how such an overlay may responsibly be shown to a physician. The acquisition audit adds that image-quality statistics alone separate the classes at AUROC 0.7482 against the model's 0.9812, a floor no prior study in this literature has measured.

The saliency-evaluation protocol-faithfulness by deletion and insertion against random-order controls, concentration by Gini and top-decile mass, class-discriminativeness by inter-group IoU against a split-half ceiling-is attributor- and model-agnostic and is released as a standalone package. The acquisition-confound audit is a three-statistic, near-zero-cost test that bounds how much of any imaging classifier's accuracy is attributable to capture conditions, and belongs in the default reporting of this literature. The embedding-geometry diagnostic identifies the separation ratio, not F1, as the criterion for selecting a triplet margin when the head reads pre-normalization features.

A control-sector ablation-retraining on an area-matched sector to which no chart assigns the lungs, and on the full annulus-determines whether the chart contributes anything, and is the condition under which the iridological reading should be withdrawn. Then a new collection with acquisition matched across classes, archiving pre-crop images, under consent permitting data sharing; multi-centre validation and prospective testing at real-world prevalence, with layer-wise learning-rate decay and a margin sweep scored on the separation ratio; and comparison against class-aware transformer attribution on the released protocol, with multimodal clinical covariates. Until at least the first two are complete, the framework is a research prototype rather than a clinical instrument, and the underlying question-whether iris imaging carries reproducible pulmonary information independent of acquisition and of the iridological premise-remains open, better posed than before.

ACKNOWLEDGMENTS

The authors extend their appreciation to Thu Dau Mot University, and Posts and Telecommunications Institute of Technology (PTIT) in Vietnam for supporting this research.

FUNDING INFORMATION

This research is funded by Thu Dau Mot University, Ho Chi Minh City, Vietnam under grant number DT.26.1-005.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Hung Ho-Dac	✓	✓	✓	✓	✓				✓	✓	✓		✓	✓
Van-Len Vo	✓				✓	✓			✓			✓		
Trong-Thua Huynh	✓					✓	✓	✓		✓		✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing-**O**riginal Draft
E : Writing-Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

INFORMED CONSENT

We have obtained informed consent from all individuals included in this study.

DECLARATION OF ETHICS

The study protocol was reviewed and approved by the institutional ethics committee (Thu Dau Mot University-TDMU).

DATA AVAILABILITY

The iris images and clinical labels cannot be shared: they are held under a data protection agreement with the collaborating centre and participants consented to research use rather than to redistribution. All code is publicly available at https://tinyurl.com/5xp7mtwt.

REFERENCES

[1] A. Simón, D. M. Worthen, and J. A. Mitas, "An evaluation of iridology," *Journal of the American Medical Association*, vol. 242, no. 13, pp. 1385–1389, 1979, doi: 10.1001/jama.1979.03300130029014.
[2] P. Knipschild, "Looking for gall bladder disease in the patient's iris," *British Medical Journal*, vol. 297, no. 6663, pp. 1578–1581, 1988, doi: 10.1136/bmj.297.6663.1578.
[3] L. Iuliano, D. Veritti, and P. Lanzetta, "Iridology: biological plausibility, clinical evidence, and implications for ophthalmic practice," *Graefe's Archive for Clinical and Experimental Ophthalmology*, vol. 264, pp. 3027–3035, 2026, doi: 10.1007/s00417-026-07342-w.
[4] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature Medicine*, vol. 25, no. 1, pp. 44–56, Jan. 2019, doi: 10.1038/s41591-018-0300-7.
[5] G. Litjens *et al*., "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, Dec. 2017, doi: 10.1016/j.media.2017.07.005.
[6] A. Bansal, R. Agarwal, and R. K. Sharma, "Pre-diagnostic tool to predict obstructive lung diseases using iris recognition system," in *Smart Innovations in Communication and Computational Sciences*, Singapore: Springer, 2019, pp. 71–79, doi: 10.1007/978-981-10-8968-8_7.
[7] M. U. Rehman, M. Driss, A. Khakimov, and S. Khalid, "Non-invasive early diagnosis of obstructive lung diseases leveraging machine learning algorithms," *Computers, Materials & Continua*, vol. 72, no. 3, pp. 5681–5697, 2022, doi: 10.32604/cmc.2022.025840.

[8] E. M. Kusumaningtyas, A. R. Barakbah, and A. A. Hermawan, "Feature extraction for application of heart abnormalities detection through iris based on mobile devices," *EMITTER International Journal of Engineering Technology*, vol 5, no. 2, pp. 312-327, 2017, doi: 10.24003/emitter.v5i2.202.
[9] T. T. Huynh, D. T. Huynh, H. P. Vu, N. M. Tuan, N. T. T. Hai, and T. L. Thanh, "ReLiris: lung damage pre-screening via iridology," *Franklin Open*, vol. 12, p. 100362, Sep. 2025, doi: 10.1016/j.fraope.2025.100362.
[10] P. Samant and R. Agarwal, "Machine learning techniques for medical diagnosis of diabetes using iris images," *Computer Methods and Programs in Biomedicine*, vol. 157, pp. 121–128, Apr. 2018, doi: 10.1016/j.cmpb.2018.01.004.
[11] F. Özbilgin and Ç. Kurnaz, "An alternative approach for determining the cholesterol level: iris analysis," *International Journal of Imaging Systems and Technology*, vol. 32, no. 4, pp. 1159–1171, Jul. 2022, doi: 10.1002/ima.22689.
[12] F. Özbilgin, Ç. Kurnaz, and E. Aydın, "Prediction of coronary artery disease using machine learning techniques with iris analysis," *Diagnostics*, vol. 13, no. 6, p. 1081, Mar. 2023, doi: 10.3390/diagnostics13061081.
[13] F. Özbilgin, Ç. Kurnaz, and E. Aydın, "Non-invasive coronary artery disease identification through the iris and bio-demographic health profile features using stacking learning," *Image and Vision Computing*, vol. 146, p. 105046, Jun. 2024, doi: 10.1016/j.imavis.2024.105046.
[14] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: transformers for image recognition at scale," *arXiv Preprint*, 2020, doi: 10.48550/arXiv.2010.11929.
[15] T. Matsuyama, H. Watanabe, and N. Takahashi, "Performance comparison of vision transformer- and CNN-based image classification using cross entropy: lung cancer discrimination from CT images," *Journal of Biomedical Science and Engineering*, vol. 17, no. 9, pp. 157–170, 2024, doi: 10.4236/jbise.2024.179012.
[16] N. Malaviya, M. Rahevar, A. Virani, A. Ganatra, and K. Bhuva, "LViT: Vision Transformer for Lung cancer Detection," in *2023 International Conference on Artificial Intelligence and Smart Communication (AISC)*, Greater Noida, India, 2023, pp. 93-98, doi: 10.1109/AISC56616.2023.10085230.
[17] J. S. Kim *et al.*, "Parallel-wise global and local attention vision transformer-based generative adversarial network using Fourier transform loss for generating fake iris image," *Engineering Applications of Artificial Intelligence*, vol. 155, p. 111080, Sep. 2025, doi: 10.1016/j.engappai.2025.111080.
[18] S. A. Latif, K. Sidek, and A. Hashim, "An efficient iris recognition technique using CNN and vision transformer," *Journal of Advanced Research in Applied Sciences and Engineering Technology*, vol. 34, no. 2, pp. 235–245, Dec. 2023, doi: 10.37934/araset.34.2.235245.
[19] 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 *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.
[20] H. Chefer, S. Gur, and L. Wolf, "Transformer Interpretability Beyond Attention Visualization," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, 2021, pp. 782-791, doi: 10.1109/CVPR46437.2021.00084.
[21] S. Abnar and W. Zuidema, "Quantifying attention flow in transformers," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 4190–4197, doi: 10.18653/v1/2020.acl-main.385.
[22] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': explaining the predictions of any classifier," in *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.
[23] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 4765–4774, 2017.
[24] S. R. J., Y. R., and V. S K., "Disease Detection Based on Iris Recognition," in *2023 International Conference on Energy, Materials and Communication Engineering (ICEMCE)*, Madurai, India, 2023, pp. 1-5, doi: 10.1109/ICEMCE57940.2023.10433979.
[25] B. Jensen, *The Science and Practice of Iridology*. Escondido, CA, USA: Bernard Jensen Enterprises, 1952.
[26] V. Petsiuk, A. Das, and K. Saenko, "RISE: randomized input sampling for explanation of black-box models *arXiv preprint*, 2018, doi: 10.48550/arXiv.1806.07421.
[27] D. Lakens, "Equivalence tests: a practical primer for t tests, correlations, and meta-analyses," *Social Psychological and Personality Science*, vol. 8, no. 4, 2017, doi: 10.1177/1948550617697177.
[28] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, 2018.
[29] N. Arun *et al*., "Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging," *Radiology: Artificial Intelligence*, vol. 3, no. 6, 2021, doi: 10.1148/ryai.2021200267.
[30] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019, doi: 10.1038/s42256-019-0048-x.

BIOGRAPHIES OF AUTHORS



Hung Ho-Dac received the engineer degree in Information Technology from Post and Telecommunication Institute of Technology (PTIT), Vietnam, in 2014, and the Master degree in Information System from in PTIT, Vietnam, in 2016. Currently, he is a lecturer/researcher Digital Transformation Center, Thu Dau Mot University, Vietnam. His research interests include blockchain, AR/VR, cloud computing, and machine learning. He can be contacted at email: hunghd@tdmu.edu.vn.



Van-Len Vo received his M.S. degree in Information Systems from Thu Dau Mot University, in 2020. Currently, he is lecturer/researcher in Institute of Digital Technology, Thu Dau Mot University, Vietnam. His research interests are cloud computing, IoT, and AI. He can be contacted at email: lenvv@tdmu.edu.vn.



Trong-Thua Huynh is currently the head of Faculty of Information Technology II at the Posts and Telecommunications Institute of Technology (PTIT), Vietnam. He received a bachelor's degree in Information Technology from Ho Chi Minh City University of Natural Sciences, a master's degree in Computer Engineering from Kyung Hee University, Korea, and a Ph.D. degree in computer science from the Ho Chi Minh City University of Technology, Vietnam National University at Ho Chi Minh City. His key areas of research include cybersecurity, AI and big data, and intelligent information systems. He can be contacted at email: thuaht@ptit.edu.vn.