Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Reliability-Aware Comparative Analysis of CNN and Swin Transformer Models for Diabetic Retinopathy Classification

Yelizaveta Kabanova

yelizavetakabanova@gmail.com

National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”

Nataliia Kuznietsova

National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”

Márta Seebauer

Obuda University

Research Article

Keywords: diabetic retinopathy, deep learning, uncertainty estimation, medical image classification, predictive reliability, explainable artificial intelligence

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10131127/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Reliability-Aware Comparative Analysis of CNN and Swin Transformer Models for Diabetic Retinopathy Classification

Yelizaveta Kabanova[1,2*], Nataliia Kuznietsova[1,2,3] and Márta Seebauer[4]

[1]Department of Mathematical Methods of System Analysis, Institute for Applied System Analysis, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine.
[2] Computer Science Department, Risk Assessment and Financial Analytics Scientific Laboratory, Kyiv School of Economics, Kyiv, Ukraine.
[3]Department of Information Technology, Faculty of Digital Technologies, EPAM, American University Kyiv, Kyiv, Ukraine.
[4]Department of Electrical and Mechatronics Engineering, Alba Regia Faculty, Obuda University, Székesfehérvár, Hungary.

*Corresponding author(s). E-mail(s): yelizavetakabanova@gmail.com;
Contributing authors: natalia-kpi@ukr.net;
seebauer.marta@amk.uni-obuda.hu;

Abstract

Background: One of the reasons people lose their eyesight is diabetic retinopathy (DR). Convolutional neural networks (CNNs) and Vision Transformers (ViTs), two deep learning techniques, have shown excellent performance in automatic DR categorisation. Nevertheless, comparative studies of these designs are frequently conducted under various experimental conditions and primarily focus on classification accuracy. Moreover, prediction reliability and uncertainty estimation are still not properly investigated.

Methods: This paper presents an uncertainty-aware comparative analysis of a DenseNet121-based CNN and a Swin Transformer architecture for DR classification under identical preprocessing, training, and evaluation conditions. Both models were evaluated on the APTOS 2019 dataset using conventional classification metrics, uncertainty-related metrics, the Reliability Score (RS) proposed by

the authors, and perturbation-based explainability metrics. Monte Carlo Dropout was applied to estimate predictive uncertainty during inference.
Results: The results demonstrated that the Swin Transformer achieved slightly higher classification performance, obtaining a Quadratic Weighted Kappa (QWK) of 0.938 and an F1-score of 0.860, while the CNN model achieved higher confidence (Mean Confidence = 0.749) and more stable and reliable prediction behavior. Explainability analysis additionally revealed differences in attention behavior between the architectures, where the CNN produced more localized explanations focused on lesion-like retinal regions, while the Transformer relied on more spatially distributed attention patterns.
Conclusions: The obtained findings demonstrate that higher classification performance does not necessarily correspond to higher predictive reliability in medical image analysis. The study highlights the importance of integrating uncertainty estimation and explainability analysis into deep learning systems for trustworthy diabetic retinopathy diagnosis.

Keywords: diabetic retinopathy, deep learning, uncertainty estimation, medical image classification, predictive reliability, explainable artificial intelligence

1 Introduction

Diabetic retinopathy (DR) is one of the causes of vision loss. It occurs when high blood sugar gradually destroys the retina's blood vessels, leading to issues like microaneurysms, haemorrhages, exudates, and the formation of new blood vessels. Images of the back of the eye show these issues, which are used to categorise the illness [1]. For prompt treatment and the avoidance of irreversible blindness, early detection and precise staging of DR are crucial.

In clinical practice, retinal examination is performed manually by ophthalmologists using specialized imaging techniques, which can be time-consuming and highly dependent on clinical expertise [2]. As a result, automated computer-aided diagnosis systems have become an important research direction in medical image analysis.

Recent advances in convolutional neural networks (CNNs) have demonstrated strong performance in diabetic retinopathy classification tasks [3]. At the same time, Vision Transformers (ViTs) have emerged as an alternative paradigm for computer vision tasks [4]. Unlike CNNs, Transformer-based architectures rely on self-attention mechanisms that capture long-range spatial dependencies and global contextual relationships within an image. Among them, the Swin Transformer has demonstrated promising results for DR classification due to its hierarchical structure and computational efficiency [5, 6].

Despite the rapid adoption of Transformer-based models in medical imaging, several important limitations remain unexplored. First, many studies focus primarily on improving classification accuracy, while considerably less attention is paid to predictive reliability and uncertainty estimation. In medical contexts, this aspect is especially important because highly confident but incorrect predictions may lead to serious clinical consequences. Second, assessing architectural differences objectively is made more

difficult by the fact that comparisons between CNN and Transformer architectures are frequently carried out under various preprocessing pipelines, augmentation procedures, or training settings.

Another limitation is the insufficient comparative analysis of attention mechanisms in CNN and Transformer architectures. While Transformer-based models rely on self-attention to capture global contextual relationships across the retinal image, CNN-based approaches typically employ localized spatial and channel attention mechanisms. However, the differences in how these attention strategies influence feature extraction, uncertainty behavior, and focus on clinically relevant retinal regions remain insufficiently explored in the context of diabetic retinopathy classification.

A review of recent studies related to Transformer-based diabetic retinopathy classification is presented in the following section.

2 Related Work

Transformer-based architectures have replaced traditional convolutional neural networks (CNNs) in automated diabetic retinopathy (DR) categorisation. Through self-attention mechanisms, Vision Transformers (ViTs) have proven capable of capturing long-range contextual dependencies. However, despite considerable progress in classification performance, important limitations remain regarding reliability, uncertainty estimation, and comparative analysis of attention mechanisms.

One of the early applications of Vision Transformers to DR classification was presented by Mutawa and Sai Sruthi [7]. The authors applied a ViT model to fundus image classification and investigated different optimization strategies, achieving an F1-score of approximately 0.73. However, it remained limited by small dataset sensitivity and did not investigate interpretability, uncertainty estimation, and attention behavior.

Wu et al. [8] further explored ViT-based DR grading by implementing six Vision Transformer variants on highly imbalanced retinal datasets. Although their work demonstrated the potential of Transformer architectures for DR severity grading, it lacked systematic reliability analysis, uncertainty estimation, and comparison with attention-based CNN models under identical experimental conditions.

Subsequent studies focused on improving locality modeling and features representation in transformers. Hossain et al. [9] proposed a Vision Transformer architecture incorporating Shifted Patch Tokenization (SPT) and Locality Self-Attention (LSA) with StyleGAN2-based augmentation and comparison against CNN transfer-learning models. Nevertheless, the comparison between CNN and Transformer approaches was performed using different preprocessing and training pipelines, making the evaluation less controlled and complicating objective conclusions regarding architectural differences.

Several studies proposed modified Transformer architectures for retinal image analysis. Zhaomin et al. [10] introduced a Swin Transformer-based framework with enhancement strategies for improving low-level retinal feature extraction. Sun et al. [11] proposed lesion-guided attention mechanisms that explicitly focus on pathological retinal regions. Although these studies improved feature localization and classification performance, they did not investigate predictive uncertainty or compare Transformer attention behavior with CNN-based attention mechanisms.

Hybrid Transformer-based approaches have also been explored. Jha et al. [12] combined Vision Transformer feature extraction with SVM classification, demonstrating competitive performance. However, the proposed architecture did not incorporate uncertainty estimation and reliability analysis.

Several recent studies emphasized the importance of uncertainty-aware DR classification. Jaskari et al. [13] investigated multiple approximate Bayesian deep learning methods across the EyePACS, Messidor-2, and APTOS datasets. More recently, Chougule [14] proposed an uncertainty-aware ensemble Vision Transformer framework with Monte Carlo Dropout–based Bayesian inference for DR classification. Although these studies addressed predictive uncertainty, they did not systematically compare CNN and Transformer architectures or analyse the relationship between their attention mechanisms and uncertainty behavior.

In our previous work [15], we proposed an uncertainty-aware CNN framework with Adaptive CBAM attention and Monte Carlo Dropout, systematically investigating the interaction between preprocessing strategies, class balancing, and predictive reliability under controlled experimental conditions. An original Reliability Score (RS) metric was additionally introduced for joint evaluation of classification quality, confidence, and uncertainty behavior.

In another study [16], we conducted a confidence-aware comparison between classical machine learning approaches based on HOG descriptors combined with SVM and Random Forest classifiers, and an uncertainty-aware CNN architecture. The findings showed that uncertainty-aware CNN models significantly outperform traditional methods in terms of prediction reliability and classification performance.

Despite these developments, a number of significant gaps are still not fully investigated. First, CNN and Transformer architectures are often compared under different preprocessing and training conditions, undermining a fair architectural comparison. Second, uncertainty-aware analyses of Vision Transformers for DR classification remain limited. Third, the influence of different attention mechanisms on predictive uncertainty, reliability, and explainability has not been systematically investigated.

To address these limitations, this study presents a comparative uncertainty-aware analysis of a CNN architecture with an Adaptive CBAM attention mechanism and a Swin Transformer model for DR classification under identical experimental conditions. In addition to conventional classification metrics, the study evaluates predictive uncertainty, confidence behavior, reliability, and explainability characteristics of both approaches, as well as the proposed integrated metric developed in the former study [15].

The aim of this study is to perform an uncertainty-aware analysis between a convolutional neural network with Adaptive CBAM attention and a Swin Transformer architecture for diabetic retinopathy classification under identical experimental conditions in order to analyse their differences and determine which approach provides more reliable and robust predictions for clinical decision support.

The main contributions of this study are summarized as follows:

- Development of an experimental framework for fair comparison of CNN and Transformer architectures under identical preprocessing, augmentation, balancing, and training conditions.

- Extension of Monte Carlo Dropout–based uncertainty estimation to the Swin Transformer architecture and comparative analysis of uncertainty behavior between convolutional and Transformer-based models.
- Demonstration that higher classification performance does not necessarily correspond to higher predictive reliability in medical image classification tasks.
- Application of the proposed Reliability Score (RS) metric for joint evaluation of predictive performance and uncertainty estimation.
- Comparative explainability analysis of CNN and Transformer attention mechanisms using perturbation-based evaluation metrics and qualitative visualizations.

3 Methods

This study follows an experimental pipeline designed to ensure a fair comparison between convolutional and Transformer-based approaches for diabetic retinopathy stage classification. Under a single experimental protocol, the workflow comprises dataset preparation, preprocessing, handling of class imbalance, model training, uncertainty estimation, and evaluation.

3.1 Dataset

The APTOS 2019 Blindness Detection (APTOS 2019 BD) dataset was used in this study [17], containing 3,662 color RGB fundus images taken at Aravind Eye Hospital in India. Each image is labeled into five classes by expert ophthalmologists, representing different stages of diabetic retinopathy severity (Figure 1):

- 0 – No DR,
- 1 – Mild DR,
- 2 – Moderate DR,
- 3 – Severe DR,
- 4 – Proliferative DR.

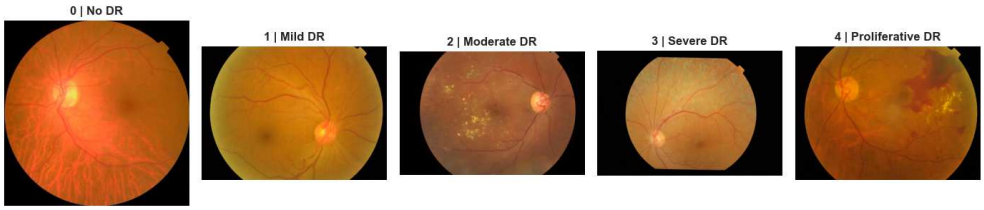


Fig. 1 Fundus images for each DR stage from the APTOS 2019 Dataset

The dataset has a number of issues, such as poor image quality, noise, fluctuations in illumination, and extreme class imbalance, because it comes from actual clinical settings. Variations in contrast, brightness, resolution, and image orientation were also found by exploratory data analysis.

Using stratified sampling based on class labels, the dataset was divided between 80% training and 20% validation sets. Model optimisation was done using the training set, and hyperparameter adjustment and performance assessment were done using the validation set.

3.2 Data preprocessing

In medical image analysis, preprocessing is essential since model performance is strongly impacted by input quality. The dataset's fundus images have artefacts and dark borders that could damage training stability as they do not contain any diagnostic information. Therefore, a preprocessing pipeline was designed to remove irrelevant regions, enhance image quality, and improve the visibility of retinal structures [15].

The preprocessing steps included cropping of black borders using intensity thresholding, resizing images to 224×224 pixels, normalization of pixel values to the [0,1] range, Gaussian smoothing for noise reduction, linear enhancement for contrast improvement, and CLAHE (Contrast Limited Adaptive Histogram Equalization) for local contrast enhancement.

These preprocessing steps improved the visibility of fine retinal structures and increased training stability. Figure 2 compares the original image and the preprocessed image.

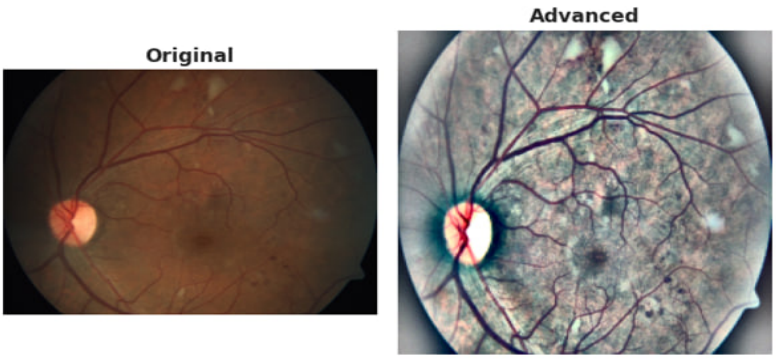


Fig. 2 Original and preprocessed image

3.3 Class imbalance handling

Class imbalance is one of the biggest challenges when working with real-world data. Modeling results are frequently skewed toward majority classes as a result of this imbalance.

Class 0 dominates with nearly 50% of all photos (1805 images), followed by class 2 with 27.3% (999 images), class 1 and class 3 with roughly 10% (370 images) and 8% (295 images), respectively, and class 4 with the lowest proportion of 5.3% (193 images).

A more balanced training distribution was produced by using oversampling-based augmentation to minority classes in order to solve this problem. This approach is commonly used in medical imaging tasks where rare cases are clinically important [17].

As a result, the total number of images increased to 9,025, with 7,220 in the training set and 1,805 in the validation set.

Additionally, Focal Loss [18] was used to further mitigate class imbalance. Unlike standard cross-entropy, Focal Loss reduces the contribution of easily classified examples and enhances the influence of complex examples.

3.4 Data augmentation

Data augmentation was applied to increase the diversity of the data set through random transformations. Two levels of augmentation were used:

- During oversampling: horizontal flipping, brightness adjustment (20%), and rotation (15°).
- During training: horizontal flipping, brightness adjustment (20%), rotation (10°), zoom (±10%), and width and height shifts (±10%).

As a result, the models were trained on a wider set of visual variations, which improved the ability to recognize pathologies in new images.

3.5 Convolutional neural network model architecture

In this study, the configuration of the best-performing CNN model developed in the preliminary experiments [15] was selected for comparison with transformer approaches.

Figure 3 illustrates the deep learning architecture for diabetic retinopathy classification.

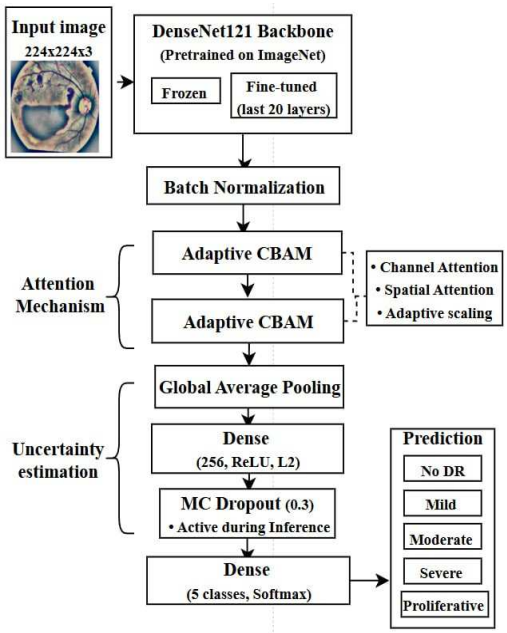


Fig. 3 Proposed deep learning architecture for diabetic retinopathy classification [16]

The DenseNet121 architecture was chosen as the backbone due to its efficiency while working with limited and unbalanced datasets and the ability to use the features of the previous layers effectively, which contributes to better model resistance to noise and artifacts in images [19].

To improve feature representation, an Adaptive Convolutional Block Attention Module (Adaptive CBAM) was integrated into the network. It combines channel and spatial attention and dynamically adjusts attention strength based on feature variance.

Global average pooling was then used to reduce feature dimensionality, followed by a fully connected layer with 256 neurons and ReLU activation. To estimate prediction uncertainty, the Monte Carlo Dropout approach [20] was used during inference.

3.6 Swin Transformer model architecture

This study used the Swin Transformer [21], a hierarchical Vision Transformer intended for effective representation learning on high-resolution pictures, as an alternative to the convolutional architecture.

The Swin Transformer creates feature representations via self-attention mechanisms, which enable the model to dynamically record interactions between spatially distant regions, in contrast to traditional CNNs that rely on fixed convolutional kernels and local receptive fields. When classifying diabetic retinopathy, this characteristic is especially crucial since diseased abnormalities may be dispersed throughout the retinal picture rather than localised in a single local location.

The model's hierarchical architecture is conceptually comparable to convolutional neural networks. The input image is first split up into 4×4 pixel non-overlapping patches, each of which is handled as a token. These tokens go through a series of Transformer blocks after being linearly projected into a higher-dimensional feature space. Neighbouring patches are joined at each level, increasing feature dimensionality while decreasing spatial resolution.

Every stage is made up of many Swin Transformer blocks, each of which has a two-layer multilayer perceptron (MLP) with GELU activation after Window-based Multi-head Self-Attention (W-MSA) and Shifted Window Multi-Head Self-Attention (SW-MSA) blocks. Each module is preceded by layer normalisation (LN), and training is stabilised using residual connections.

Window-based self-attention, which limits attention computation to local regions of size $M \times M$, is a crucial part of the model. Compared to global attention, this greatly lowers computational complexity. For a fixed M, the complexity scales linearly with image size, making the model suitable for high-resolution medical imaging tasks.

To compensate for the locality of window-based attention, the Swin Transformer introduces a shifted window partitioning strategy, where the window configuration is shifted between consecutive layers. This enables interaction between neighboring regions without increasing computational cost and improves the model's ability to capture long-range dependencies.

Self-attention is computed as:

$$\text{Attention}(Q, K, V) = \text{SoftMax}\left(\frac{QK^T}{d} + B\right)V, \tag{1}$$

where $Q, K, V \in \mathbb{R}^{M^2 \times d}$ are the query, key, and value matrices, respectively; d is the feature dimension, and B represents a learnable relative positional bias that improves spatial modeling [22].

In this study, the Swin Transformer was implemented using the following parameters:

- Patch size: 4×4;
- Window size: 7;
- Embedding dimension: 96;
- Layer numbers: {2, 2, 6, 2}.

An important aspect of this study is the comparison of attention mechanisms. The Adaptive CBAM module, which applies spatial and channel attention based on feature variance, explicitly introduces attention into the CNN model. The Swin Transformer, on the other hand, uses self-attention matrices, which implicitly use attention and dynamically compute feature interactions based on input data. This allows modeling of long-range dependencies that are difficult to capture using convolutional operations alone [21].

Therefore, the Swin Transformer serves as a fundamentally different modelling paradigm within the unified experimental framework, enabling a controlled comparison of feature extraction, attention behaviour, and predictive reliability in DR classification.

3.7 Uncertainty estimation via Monte Carlo Dropout

To evaluate prediction reliability, Monte Carlo Dropout [20] was applied during inference for both models. This method produces a set of stochastic predictions for the same input image by enabling the models to repeatedly make the prediction using the active Dropout layer. The final output is then determined by averaging the forecasts.

Based on the obtained stochastic predictions, the average values of class probabilities, prediction variance, entropy, and average model confidence were computed. This allows the analysis of the stability and reliability of the model's decisions under conditions of uncertain or noisy input data.

3.8 Experimental setup

To ensure a fair comparison between convolutional and Transformer-based approaches, both models were trained and evaluated under closely aligned experimental conditions. The hyperparameters and training strategies were selected to minimize external influences on model performance and enable systematic and controlled comparison of the architectures under consistent experimental settings. The detailed hyperparameter configuration is summarized in Table 1.

3.9 Evaluation metrics

To evaluate model performance, both standard classification [22] and uncertainty metrics were used.

Table 1 Hyperparameters for CNN and Swin Transformer models

Parameter	CNN	Swin Transformer
Backbone	DenseNet121 (ImageNet)	Swin-Tiny (ImageNet)
Input image size	224×224	224×224
Fine-tuning strategy	Unfreeze last 20 layers	Full fine-tuning
Attention mechanism	Adaptive CBAM	Window-based self-attention
Normalization	Batch Normalization	Layer Normalization
Dropout ratio	0.3	0.2
Loss function	Focal Loss	Focal Loss
Optimizer	Adam	Adam
Learning rate (Stage 1)	1×10^{-4}	1×10^{-4}
Learning rate (Stage 2)	1×10^{-5}	1×10^{-5}
Batch size	32	32
Number of epochs	30	30
Uncertainty estimation	MC Dropout (25 passes)	MC Dropout (25 passes)
Random seed	Fixed	Fixed

Accuracy – proportion of correct predictions:

$$\text{ACC} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{2}$$

Precision – proportion of correctly identified positive cases:

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{3}$$

Recall – the ability of the model to find all positive cases:

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{4}$$

F1-score – a harmonic mean of precision and recall:

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{5}$$

where TP, TN, FP, FN stand for true positive, true negative, false positive, and false negative values, respectively.

Quadratic Weighted Kappa (QWK) – takes into account the ordinal nature of the classes and different error weights between adjacent stages of the disease:

$$\text{QWK} = 1 - \frac{\sum_{i,j} w_{ij} O_{ij}}{\sum_{i,j} w_{ij} E_{ij}}, \tag{6}$$

with

$$w_{ij} = \frac{(i-j)^2}{(C-1)^2}, \tag{7}$$

where O_{ij} is the observed confusion matrix, E_{ij} is the expected matrix under random agreement, and w_{ij} is the quadratic weight.

Area Under the ROC Curve (AUC) – the trade-off between true positive rate and false positive rate across different classification thresholds.

Classical metrics can provide classification accuracy, but are unable to assess how confident the model is in making a prediction. Uncertainty-based metrics were computed using Monte Carlo Dropout, including mean confidence, entropy, and variance.

Confidence – the model's average confidence, calculated as the maximum predicted class probability after averaging the Monte Carlo Dropout predictions.

Entropy – a measure of predictive uncertainty:

$$\mathrm{H}(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i) \tag{8}$$

Variance – a measure of the dispersion of prediction values:

$$\mathrm{Var}(p) = \frac{1}{n} \sum_{i=1}^{n} (p_i - \bar{p})^2 \tag{9}$$

where p_i is the probability of class i and $\bar{p}_i$ is the mean value of all predictions.

In addition, the authors' integrated metric Reliability Score (RS) [7] was introduced to jointly evaluate predictive performance and uncertainty:

$$\mathrm{RS} = \alpha \ln(\mathrm{QWK} + \varepsilon) + \beta(1 - H_{\mathrm{norm}} + \varepsilon) + \gamma \ln(C + \varepsilon), \tag{10}$$

where H_{norm} is the normalized mean entropy, C is the mean prediction confidence, $\alpha = 1, \beta = 0.5, \gamma = 0.5$ are weighting coefficients, and $\varepsilon > 0$ is a small constant. RS does not have a fixed range, as it depends on the logarithmic components.

However, higher RS values correspond to more reliable and confident model predictions.

3.10 Explainability analysis

In medical image analysis, even minor details can contain important diagnostic information. Therefore, explainability methods are essential not only for improving model transparency but also for assessing whether the model relies on clinically meaningful regions. In this study, both qualitative and quantitative explainability approaches were applied to the CNN and Swin Transformer models.

Pixel-level contributions to the CNN model's prediction were visualised using Saliency Maps. In order to estimate how sensitive the prediction is to changes in each pixel, this approach calculates the gradient of the output class score with respect to the input image [23].

Grad-CAM was used to create class-discriminative localisation maps for the Swin Transformer architecture. Grad-CAM utilizes gradients of the target class flowing into

the last attention-based feature maps to produce a coarse localization of important regions [24].

These visualization techniques provide qualitative insight into model behavior, but do not allow for systematic comparison between models. To enable quantitative evaluation of explainability, three perturbation-based metrics were used: Deletion AUC, Insertion AUC, and Energy Concentration [25].

Deletion AUC evaluates how rapidly the model confidence decreases when the most important regions are progressively removed from the input. A lower area under the curve indicates stronger reliance on the identified salient regions [26].

Insertion AUC measures the opposite: the model is evaluated while progressively adding the most important regions to an initially blank or occluded image. Higher insertion AUC indicates that the selected regions are sufficient to recover model confidence [26].

Energy Concentration quantifies how much of the total explanation mass is concentrated in the most relevant regions. Specifically, the normalized explanation map is sorted, and the proportion of total energy contained in the top-k percentile regions is computed. Higher values indicate more spatially concentrated explanations.

These methods are essential for comparing CNN and Transformer architectures in terms of whether their decision-making processes align with clinically meaningful image regions.

4 Results

This section presents the experimental results of the proposed CNN and Swin Transformer models evaluated under identical training and preprocessing conditions. The study focuses on model accuracy, uncertainty estimations, and classification performance in the detection of diabetic retinopathy. The overall quantitative results are summarised in Table 2.

Findings show that both models performed well on all evaluation metrics, demonstrating the effectiveness of both methods for classifying diabetic retinopathy. On the other hand, uncertainty estimation and prediction reliability clearly differ.

With a QWK of 0.938, an accuracy of 0.860, and a F1-score of 0.860, the Swin Transformer had the greatest overall classification performance. This may be related to the ability of self-attention mechanisms to model complex spatial relationships across the retinal image.

In terms of uncertainty estimation, the Swin Transformer produced higher average entropy (0.741) and predictive variance (0.024) compared to the CNN model, indicating more distributed probability outputs and less concentrated confidence distributions. At the same time, the model maintained a relatively strong mean confidence of 0.704.

The entropy distribution (Figure 4) did not show extreme skewness, suggesting that the model does not produce overconfident predictions in uncertain cases.

Table 2 Classification results for the CNN and Swin Transformer models

Model	QWK	ACC	F1	AUC	RS	Uncertainty	Mean Entropy	Mean Confidence
Swin Transformer	0.938	0.860	0.860	0.977	0.578	0.024	0.741	0.704
CNN DenseNet121	0.922	0.856	0.855	0.979	0.631	0.004	0.645	0.749

Note: QWK – Quadratic Weighted Kappa; ACC – Accuracy; F1 – F1-score; AUC – Area Under the ROC Curve; RS – Reliability Score. Higher values of QWK, ACC, F1, AUC, and Mean Confidence indicate better model performance, whereas lower values of Uncertainty and Mean Entropy indicate more reliable predictions.

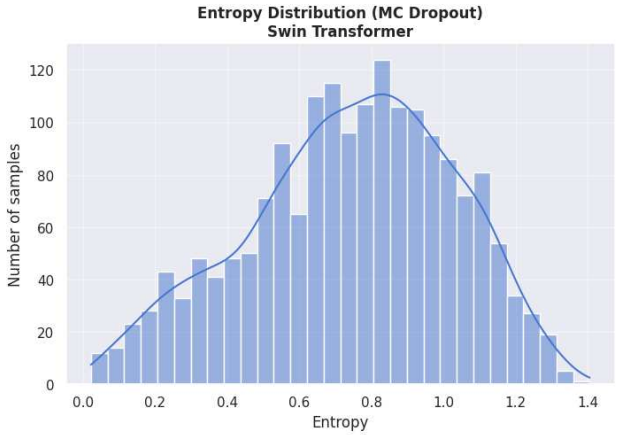


Fig. 4 The Swin Transformer model predicted entropy distributions

The reliability of predictions was further evaluated using the proposed Reliability Score (RS), which combines classification performance and uncertainty estimation into a unified metric. The Swin Transformer achieved an RS of 0.578, compared to 0.631 for the CNN model.

In comparison, the CNN model demonstrated slightly lower classification performance, with a QWK of 0.922, an accuracy of 0.856, and an F1-score of 0.855. However, it showed higher mean confidence (0.749) and lower entropy (0.645), indicating more confident and less uncertain predictions. Furthermore, CNN's higher RS value of 0.631 indicates that, in the context of the uncertainty-based evaluation methodology, its predictions are more stable.

Additionally, the entropy distribution was spread out and generally tended toward smaller uncertainty values (Figure 5).

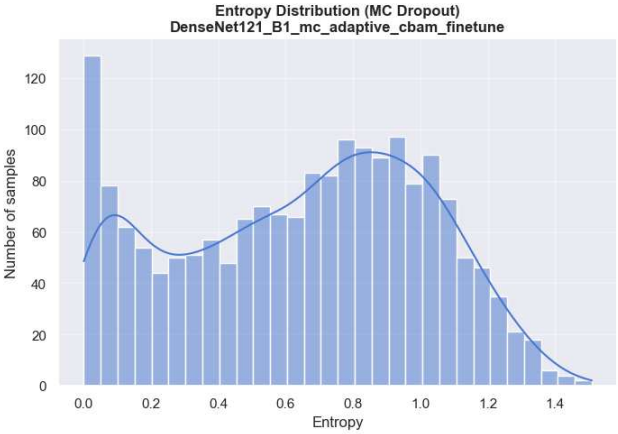


Fig. 5 The CNN model predicted entropy distributions

Confusion matrix analysis showed that the CNN model performed slightly better in most classes, especially classes 0, 1, 2, and 4, whereas the Swin Transformer performed the best in class 3. Similar misclassification patterns were shown by both models, with class 2 showing the lowest scores (0.74 for CNN and 0.70 for Swin). The confusion matrices are depicted in Figure 6.

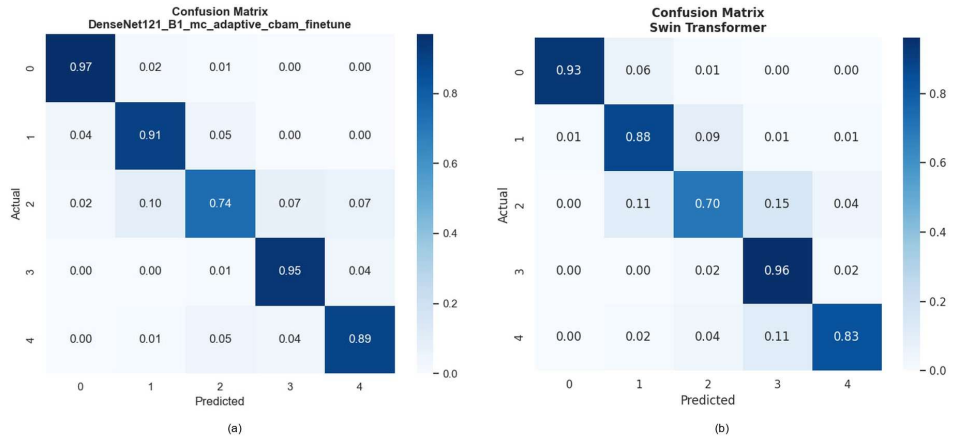


Fig. 6 Confusion matrices of (a) the CNN and (b) the Swin Transformer models

The uncertainty behaviour was further analysed through confidence and entropy distributions for correct and incorrect predictions.

For the Swin Transformer, confidence distributions (Figure 7) exhibited substantial overlap between correct and incorrect predictions, with most predictions concentrated within moderate confidence ranges (approximately 0.2-0.5). Similarly, entropy values remained relatively high across both correct and incorrect predictions (around 1.2-1.6).

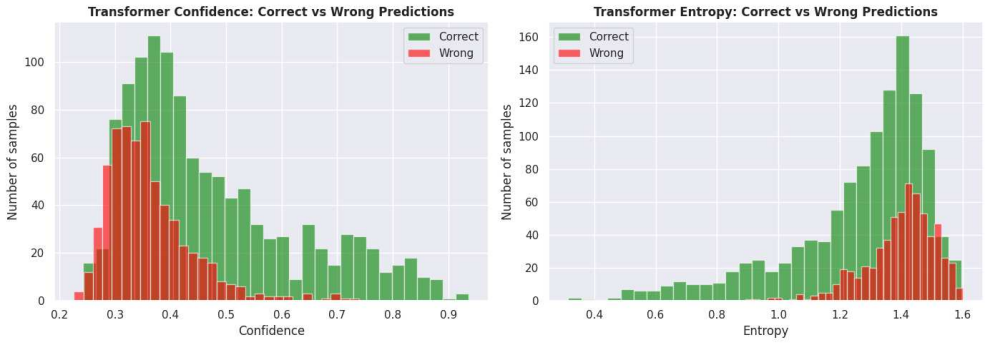


Fig. 7 Confidence and entropy distributions for correct and incorrect predictions of the Swin Transformer

In contrast, the CNN model demonstrated significantly sharper confidence distributions (Figure 8). Correct predictions were strongly concentrated near high confidence values, while incorrect predictions were associated with noticeably lower confidence. A similar trend was observed for entropy, where correctly classified samples exhibited substantially lower entropy compared to incorrect predictions. This indicates a clearer separation between confident and uncertain predictions in the CNN model.

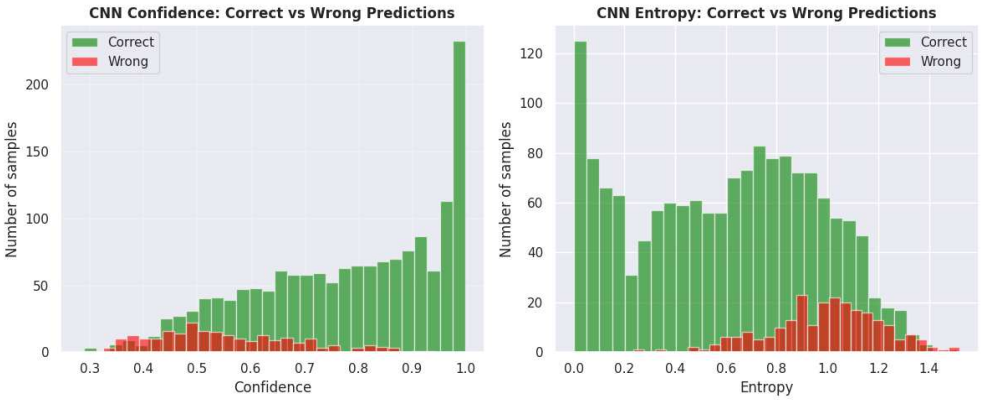


Fig. 8 Confidence and entropy distributions for correct and incorrect predictions of the CNN model

To further evaluate model interpretability, perturbation-based explanation metrics were applied. The results are summarized in Table 3.

Table 3 Results of perturbation-based metrics

Model	Deletion AUC	Insertion AUC	Energy Concentration
Swin Transformer	0.3609	0.4326	0.3690
CNN DenseNet121	0.7769	0.7801	0.3878

The Swin Transformer demonstrated significantly lower deletion AUC values, indicating that model confidence rapidly deteriorated when salient regions were removed. This suggests stronger dependence on globally distributed attention regions identified by Grad-CAM explanations.

At the same time, the CNN model achieved substantially higher insertion AUC values and slightly higher energy concentration. This indicates that the CNN explanations remain more spatially concentrated and that prediction confidence can be restored using a relatively compact subset of salient retinal regions. The higher energy concentration additionally suggests that convolutional attention mechanisms focus more locally on clinically relevant pathological structures.

Figure 9 compares the classification results for the same retinal image produced by the CNN and Swin Transformer models.

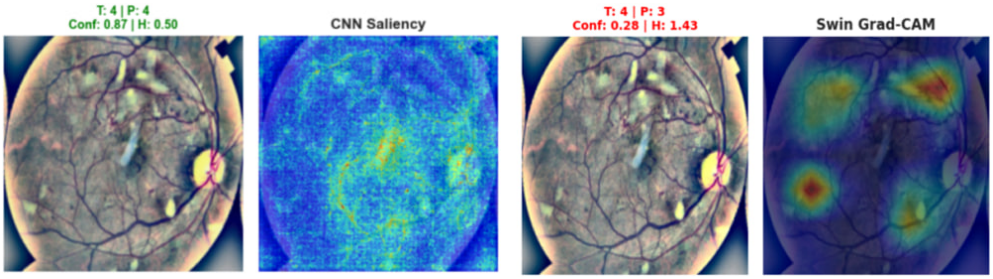


Fig. 9 Comparison of classification results for the same retinal image produced by the CNN and Swin Transformer models, including predicted confidence and entropy scores

As shown in the example above, the Swin Transformer model incorrectly classified the image as class 3 with a confidence of 0.28, while the CNN model correctly classified the same image with a confidence of 0.87. Analysis of the highlighted retinal regions demonstrated that the CNN attention mechanism focused on compact pathological structures, whereas the Transformer attention remained more spatially distributed across the retinal image.

Overall, the obtained results demonstrate that Transformer-based architectures provide slightly stronger classification performance, while CNN-based models exhibit more stable confidence behaviour, lower predictive uncertainty, and higher reliability under uncertainty-aware evaluation conditions.

5 Discussion

The results of this study demonstrate that both the Swin Transformer and the CNN model based on DenseNet121 achieved strong classification performance for diabetic retinopathy under an identical experimental setup. Despite the relatively small difference in overall classification metrics, the two architectures showed substantially different behaviour in terms of uncertainty estimation, confidence distributions, and explainability patterns.

The Swin Transformer achieved the highest QWK, accuracy, and an F1-score, but the uncertainty analysis revealed that the Swin Transformer produced noticeably higher predictive entropy and variance compared to the CNN model. The confidence distributions also demonstrated that correct and incorrect predictions remained partially overlapping, while entropy values stayed relatively high even for correctly classified samples. This behaviour suggests that the Transformer produces softer probability distributions and avoids highly overconfident predictions.

In contrast, the CNN model demonstrated more concentrated confidence distributions, lower entropy, and substantially lower predictive variance. Such clearer separation indicates more stable prediction behaviour and stronger confidence calibration within the proposed uncertainty estimation framework.

An important finding of this study is that higher classification performance does not necessarily correspond to higher predictive reliability. Although the Swin Transformer achieved stronger classification metrics, the CNN obtained a higher Reliability Score due to lower uncertainty and higher confidence in predictions. This highlights the importance of uncertainty estimation when working with medical data.

The explainability analysis also demonstrated a clear difference between the custom convolutional and Transformer-based attention mechanisms. The Swin Transformer demonstrated lower deletion AUC values, indicating stronger dependence on globally distributed image regions. In contrast, the CNN demonstrated higher energy concentration values, suggesting more localized explanations focused on retinal lesion regions.

Another key finding is that the two architectures rely on fundamentally different feature attribution strategies for retinal pathology assessment. The CNN primarily relies on localized pathological structures, while the Transformer integrates broader contextual information across the retinal image. As a result, Transformer models may provide stronger global feature representation, whereas CNN-based approaches may offer more stable confidence behaviour.

Overall, the findings support the idea of the importance of uncertainty-aware evaluation in medical image classification tasks. The study also demonstrates that attention mechanisms influence not only predictive performance but also confidence behaviour, uncertainty estimation, and explainability characteristics of models for diabetic retinopathy classification.

6 Conclusions

The aim of this study was to present an uncertainty-aware experimental framework for a fair comparison between a CNN architecture and a Swin Transformer model for diabetic retinopathy classification under identical preprocessing, augmentation, balancing, training, and evaluation conditions.

To achieve this objective, a DenseNet121-based CNN with Adaptive CBAM attention architecture and a Swin Transformer model were implemented and evaluated within a unified experimental pipeline using the APTOS 2019 dataset. Both models incorporated Monte Carlo Dropout for uncertainty estimation and were evaluated using conventional classification metrics, uncertainty metrics, the proposed Reliability Score (RS), and perturbation-based explainability metrics.

The results showed that both models achieved strong diabetic retinopathy classification performance. With a QWK of 0.938, an accuracy of 0.860, and an F1-score of 0.860, the Swin Transformer achieved slightly higher classification scores. However, the CNN model demonstrated lower predictive entropy and variance, higher mean confidence, and a higher Reliability Score, indicating more stable and reliable prediction behaviour under uncertainty-aware evaluation conditions.

An important finding of the study is that higher classification performance does not necessarily correspond to higher predictive reliability. The results additionally demonstrated substantial differences between convolutional and Transformer-based attention mechanisms. The CNN attention mechanism produced more localized and concentrated explanations focused on retinal lesion regions, while the Swin Transformer relied on more spatially distributed global attention patterns.

The obtained findings confirm the importance of integrating uncertainty estimation and explainability analysis into medical image classification systems. They also support the usefulness of integrated evaluation metrics such as the proposed Reliability Score, which enables joint assessment of classification performance, predictive confidence, and uncertainty estimation within a unified reliability-oriented framework.

The limitations of the study include the use of a single retinal imaging dataset, which may limit model generalization across different clinical environments. In addition, only one CNN and one Transformer architecture were investigated. Future work may include evaluation on larger multi-center datasets, investigation of additional Transformer variants, and hybrid architectures.

Abbreviations

DR: Diabetic Retinopathy; CNN: Convolutional Neural Network; ViT: Vision Transformer; CBAM: Convolutional Block Attention Module; GAP: Global Average Pooling; MLP: Multi-Layer Perceptron; W-MSA: Window-based Multi-Head Self-Attention; SW-MSA: Shifted Window Multi-Head Self-Attention; LN: Layer Normalization; MC Dropout: Monte Carlo Dropout; ACC: Accuracy; QWK: Quadratic Weighted Kappa; AUC: Area Under the ROC Curve; RS: Reliability Score; CLAHE: Contrast Limited Adaptive Histogram Equalization; RGB: Red-Green-Blue; ReLU: Rectified Linear Unit; GELU: Gaussian Error Linear Unit; TP: True Positive; TN: True Negative; FP: False Positive; FN: False Negative.

Declarations

Ethics approval and consent to participate.

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Consent for publication.

Not applicable.

Funding

This research received no external funding.

Availability of data and materials

The dataset used and analysed during the current study is publicly available in the APTOS 2019 Blindness Detection repository at https://www.kaggle.com/c/aptos2019-blindness-detection. The source code used for data preprocessing, model training, uncertainty estimation, and evaluation is publicly available at https://github.com/kabanovaelizaveta/diabetic-retinopathy-transformer-vs--nn.

Authors' contributions

Conceptualization, Y.K., N.K., and M.S.; methodology, Y.K., N.K.; software, Y.K.; validation, Y.K., N.K.; formal analysis, Y.K.; investigation, Y.K.; resources, Y.K.; data curation, Y.K.; writing—original draft preparation, Y.K.; writing—review and editing, Y.K., N.K., M.S.; visualization, Y.K.; supervision, N.K., M.S.; project administration, N.K., M.S. All authors have read and agreed to the published version of the manuscript.

Acknowledgements

The authors would like to thank the developers and contributors of the APTOS 2019 Blindness Detection dataset for making the retinal image data publicly available for research purposes.

References

[1] Alyoubi, W.L., Abulkhair, M.F., Shalash, W.M.: Diabetic retinopathy fundus image classification and lesions localization system using deep learning. Sensors **21**(11), 3704 (2021) https://doi.org/10.3390/s21113704

[2] Clinical Strategies in the Management of Diabetic Retinopathy. Springer, Cham (2019). https://doi.org/10.1007/978-3-319-96157-6

[3] Alqahtani, A.S., Alshareef, W.M., Aljadani, H.T., Hawsawi, W.O., Shaheen, M.H.: The efficacy of artificial intelligence in diabetic retinopathy screening: a systematic review and meta-analysis. International Journal of Retina and Vitreous **11**(1) (2025) https://doi.org/10.1186/s40942-025-00670-9

[4] Wang, Y., Deng, Y., Zheng, Y., Chattopadhyay, P., Wang, L.: Vision transformers for image classification: A comparative survey. Technologies **13**(1), 32–32 (2025) https://doi.org/10.3390/technologies13010032

[5] Mallampalli, P., R, M.: Vision transformer-based swin transformer for automated diabetic retinopathy severity detection and classification. In: 2025 Global Conference on Information Technology and Communication Networks (GITCON), pp. 1–7 (2025). https://doi.org/10.1109/gitcon65266.2025.11376796

[6] Boulaabi, M., Gader, T.B.A., Echi, A.K., Bouraoui, Z.: Enhancing dr classification with swin transformer and shifted window attention. In: Lecture Notes in

Computer Science, pp. 57–61 (2025). https://doi.org/10.1007/978-3-031-95841-0_11

[7] Mutawa, A.M., Sruthi, S.: Diabetic retinopathy classification using vision transformer. In: 2022 6th European Conference on Electrical Engineering & Computer Science (ELECS), Bern, Switzerland, pp. 25–30 (2022). https://doi.org/10.1109/elecs55825.2022.00012

[8] Wu, J., Hu, R., Xiao, Z., Chen, J., Liu, J.: Vision transformer-based recognition of diabetic retinopathy grade. Medical Physics **48**(12), 7850–7863 (2021) https://doi.org/10.1002/mp.15312

[9] Hossain, S., Chakrabarty, A., Alam, G.R.: Diabetic Retinopathy Classification using VisionTransformer Architectures and Deep Learning. Research Square (2023). https://doi.org/10.21203/rs.3.rs-3261198/v1

[10] Yao, Z., Yuan, Y., Shi, Z., Mao, W., Zhu, G., Zhang, G., Wang, Z.: Funswin: A deep learning method to analysis diabetic retinopathy grade and macular edema risk based on fundus images. Frontiers in Physiology **13**, 961386 (2022) https://doi.org/10.3389/fphys.2022.961386

[11] Sun, R., Li, Y., Zhang, T., Mao, Z., Wu, F., Zhang, Y.: Lesion-aware transformers for diabetic retinopathy grading. (2021). https://doi.org/10.1109/cvpr46437.2021.01079

[12] Jha, S., Luhach, V., Poddar, R.: Retinal malady classification using AI: A novel ViT-SVM combination architecture (2026). https://doi.org/10.48550/arXiv.2603.29181

[13] Jaskari, J., Sahlsten, J., Damoulas, T., Knoblauch, J., Sarkka, S., Karkkainen, L., Hietala, K., Kaski, K.K.: Uncertainty-aware deep learning methods for robust diabetic retinopathy classification. IEEE Access **10**, 76669–76681 (2022) https://doi.org/10.1109/access.2022.3192024

[14] Chougule, P.: Uncertainty-Aware Ensemble Vision Transformers for Ordinal Diabetic Retinopathy Grading (2026). https://doi.org/10.2139/ssrn.6542898

[15] Kabanova, Y.I., Kuznietsova, N.V., Ivanko, K.O., Kulkarni, V.: Diabetic retinopathy diagnosis using deep neural networks. Applied Aspects of Information Technology **9**(2), 220–233 (2026) https://doi.org/10.15276/aait.09.2026.16

[16] Kabanova, Y., Kuznietsova, N., Ivanko, K., Kulkarni, V.: Comparison of hog-based classical classifiers and cnn models for diabetic retinopathy diagnosis. In: Proceedings of the Computer Modeling and Intelligent Systems Conference (CMIS 2026). CEUR Workshop Proceedings, ??? (2026). In press

[17] Altalhan, M., Algarni, A., Turki-Hadj Alouane, M.: Imbalanced data problem in

machine learning: A review. IEEE Access **13**, 13686–13699 (2025) https://doi.org/10.1109/access.2025.3531662

[18] Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollar, P.: Focal loss for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1–1 (2018) https://doi.org/10.1109/tpami.2018.2858826

[19] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2261–2269 (2017). https://doi.org/10.1109/cvpr.2017.243

[20] Gal, Y., Ghahramani, Z.: Dropout as a Bayesian Approximation: Appendix (2015). https://doi.org/10.48550/arxiv.1506.02157

[21] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9992–10002 (2021). https://doi.org/10.1109/ICCV48922.2021.00986

[22] Grandini, M., Bagli, E., Visani, G.: Metrics for Multi-Class Classification: an Overview (2020). https://doi.org/10.48550/arxiv.2008.05756

[23] Simonyan, K., Vedaldi, A., Zisserman, A.: Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps (2013). https://doi.org/10.48550/arXiv.1312.6034

[24] V, A.C., B, S.K., Pradeep, S., Suraksha, P., Lin, M.: Leveraging compact convolutional transformers for enhanced covid-19 detection in chest x-rays: a grad-cam visualization approach. Frontiers in Big Data **7** (2024) https://doi.org/10.3389/fdata.2024.1489020

[25] Samek, W., Binder, A., Montavon, G., Lapuschkin, S., Muller, K.-R.: Evaluating the visualization of what a deep neural network has learned. IEEE Transactions on Neural Networks and Learning Systems **28**(11), 2660–2673 (2017) https://doi.org/10.1109/tnnls.2016.2599820

[26] Hsieh, C.-Y., Yeh, C.-K., Liu, X., Ravikumar, P., Kim, S., Kumar, S., Hsieh, C.-J.: Evaluations and Methods for Explanation through Robustness Analysis (2021). https://doi.org/10.48550/arXiv.2006.00442