Journal of Computer Science and Cybernetics, V.41, N.4 (2025), 337–354
DOI no. 10.15625/1813-9663/22005

FUSIONNETX: A HIGHLY EFFECTIVE MULTIMODAL FRAMEWORK FOR SKIN CANCER DETECTION

LAM HUNG NGUYEN[1], THANG CAP[1], HUONG BUI[2], TUONG LE[2,*]

[1] *University of Information Technology, Vietnam National University Ho Chi Minh City, Quarter 34, Linh Xuan Ward, Ho Chi Minh City, Viet Nam*
[2] *Faculty of Information Technology, HUTECH University, 475A Dien Bien Phu, Thanh My Tay Ward, Ho Chi Minh City, Viet Nam*

Abstract. Early detection of skin cancer significantly improves patient outcomes by allowing for timely intervention. This study introduces the FusionNetX framework, which is a robust multimodal model using both image data and metadata for skin cancer detection, leveraging the ISIC 2024 dataset. Our approach integrates convolutional neural networks (CNNs) and Transformer-based models to extract features from single-lesion images cropped from 3D Total Body Photographs (3D-TBP). These images, resembling close-up smartphone photos, are integrated with metadata analyzed using tree-based classifiers to enhance diagnostic accuracy. To address the extreme class imbalance in the dataset, we employ advanced sampling techniques and use stratified group cross-validation to ensure our model generalizes well across diverse patient groups. Our model demonstrates competitive performance, achieving a partial area under the Receiver Operating Characteristic curve (pAUC) of 0.18380 on cross-validation and securing the top rank with a private score of 0.17295 on the private test set. This top-ranking performance highlights the model's ability to maintain high true positive rates (TPR $\geq$ 80%) while outperforming all other teams based on private scores. Furthermore, it demonstrated strong generalization across external datasets such as HAM10000 and PH2, showcasing its robustness in detecting various skin lesion types. These results underscore the effectiveness of our multimodal approach, offering a promising solution for enhancing early skin cancer detection and improving patient outcomes.

Keywords. Skin cancer detection, ISIC 2024 dataset, CNN and transformer integration, multimodal model.

1. INTRODUCTION

The increase in skin cancer cases, particularly melanoma, has raised the demand for early and accurate diagnosis. Due to the highly invasive nature of melanoma, early detection and differentiation between malignant and benign lesions can save many lives. However, traditional diagnostic methods are becoming increasingly inadequate because of limitations in access to dermatologists, especially in underserved areas. This situation highlights the

*Corresponding author.
E-mail addresses: 22520968@gm.uit.edu.vn (L.H. Nguyen), thangcpd@uit.edu.vn (T. Cap), bd.huong@hutech.edu.vn (H. Bui), lc.tuong@hutech.edu.vn (T. Le).

urgent need for automated diagnostic tools that are both accurate and accessible. Recent advancements in machine learning, particularly with technologies such as deep neural networks [1, 2, 3] and their applications in medical diagnosis [4], have opened new possibilities for more effective detection and classification of skin lesions. The application of machine learning in medical diagnosis can not only improve accessibility and provide consistent and reliable classifications but also play a crucial role in the early detection of skin cancer. Ultimately, these advances have the potential to enhance survival rates.

The International Skin Imaging Collaboration (ISIC) has been at the forefront of providing skin cancer detection datasets, named the ISIC 2024 dataset, which comprises 401,059 images and offers a rich resource for developing deep learning approaches to distinguish between malignant and benign skin lesions. Notably, the quality of single-lesion crops obtained from 3D-TBP is comparable to that of smartphone photographs, which are regularly submitted for telehealth purposes. These captured images using 3D-TBP technology provide a holistic view of the skin surface, thereby enhancing the detection capabilities of AI models. By leveraging this extensive dataset, researchers and developers can more effectively train and validate algorithms, thereby improving diagnostic accuracy and contributing to better patient outcomes in skin cancer treatment. Collaborating and sharing such valuable resources represents a significant advance in the ongoing fight against skin cancer. However, the ISIC 2024 dataset presents a significant challenge in the form of class imbalance, with only 393 images labeled as malignant out of 401,059 samples. Additionally, integrating multimodal data by combining image features with patient metadata remains a complex task that requires sophisticated modelling techniques to achieve high performance.

To address the challenges of the ISIC 2024 dataset, this study proposes FusionNetX, a multimodal framework combining deep learning for image analysis with tree-based ensemble classifiers for metadata. The main contributions are as follows: (1) Advanced preprocessing, feature engineering, and ensemble learning to mitigate class imbalance and improve model generalization; (2) Integration of CNN and Transformer-based models for precise visual feature extraction, complemented by metadata-based tree ensembles for enhanced diagnostic accuracy; (3) Incorporation of Generalized Mean Pooling (GeM) and Adaptive Average Pooling to retain discriminative features and ensure consistent representation; And (4) demonstration of strong generalization on external datasets (HAM10000 [5], PH2 [6]), confirming robustness and reliability for clinical deployment.

The rest of this paper is organized as follows: Section 2 reviews related work on ISIC datasets, skin-cancer detection, and multimodal fusion; Section 3 describes the proposed FusionNetX method, including preprocessing and architecture; Section 4 presents and discusses the experimental results; Section 5 concludes and outlines future work.

2. RELATED WORK

2.1. ISIC 2024 dataset

The international skin imaging collaboration (ISIC) has been instrumental in advancing research in skin lesion analysis by providing comprehensive, publicly available datasets. The ISIC 2024 dataset [7, 8] builds upon these efforts, offering a more extensive and diverse collection of images captured using 3D-TBP technology, thereby enabling the development of more robust and generalizable models.

Origin and composition. The ISIC 2024 dataset, also known as SLICE-3D, was created to address the shortcomings of previous skin lesion datasets, which often relied on dermoscopic images prone to selection bias and were primarily suitable for specialist use. By utilizing 3D Total Body Photography (3D TBP) with 92 DSLR cameras [9], SLICE-3D systematically captures the entire skin surface, thereby reducing bias and producing images with optical resolution comparable to those of smartphone photos. This approach aims to facilitate the development of skin cancer screening tools applicable in broader settings like primary care and teledermatology. Data collection involved over 1,000 patients from seven international dermatology centers between 2015 and 2024, under strict ethical approval in this study.

The curation process involved the VECTRA WB360 system and standardized tools to extract over 400,000 de-identified 15×15 mm image crops with associated metadata [9]. Lesions were labeled based on pathology reports (strong labels, confirmed by biopsy [9, 10, 11]) or clinical evaluation (weak labels), resulting in a dataset with significant class imbalance favoring benign cases.

Data analysis. The ISIC 2024 dataset comprises a vast collection of 401,059 images, of which only 393 are labelled as malignant, and the remaining are benign. This extreme class imbalance poses significant challenges for machine learning models, as they must accurately identify rare malignant cases within an overwhelmingly benign dataset. Traditional models may overfit benign samples, leading to high accuracy in general but low sensitivity to malignant cases. Addressing this imbalance is crucial for achieving reliable performance in practical clinical applications where accurate malignant detection is essential.

The ISIC 2024 dataset features predominantly low-resolution images, mostly with dimensions under 256 × 256 pixels (often 128 × 128–156 × 156), which is significantly lower than those in datasets like ISIC 2020. While this presents a challenge for model accuracy in capturing subtle features, the images still retain enough detail to analyze crucial characteristics for malignancy detection, such as irregular borders, color variation, and asymmetry. Furthermore, image quality varies significantly depending on the imaging source and settings. Clinical images (from dermoscopes/3D systems) are typically of higher quality, providing valuable information such as the "ugly duckling sign." In contrast, telemedicine images taken by patients on smartphones often suffer from lower quality due to factors like poor lighting and inconsistent focus.

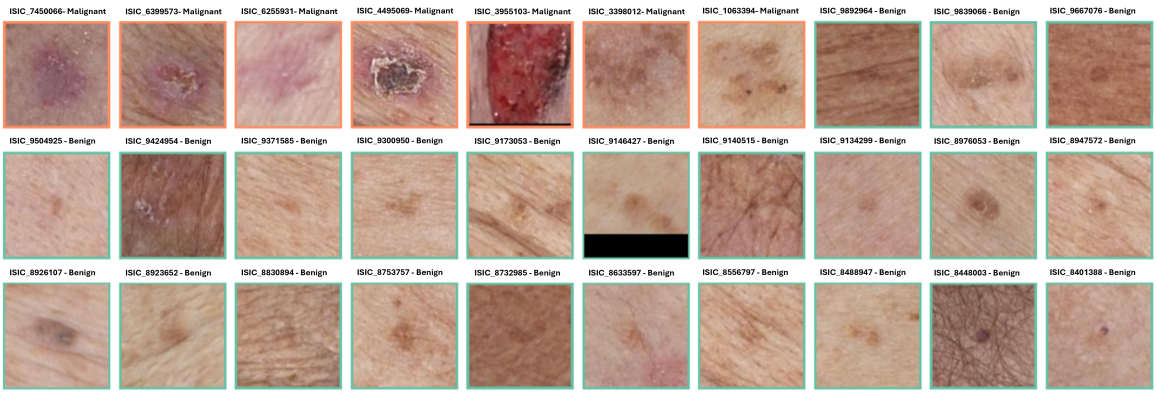

Figure 1: Sample images from a patient (30 out of 209 images)

The dataset contains diverse benign and malignant lesions from various patients, showcasing significant variation in appearance (texture, color and shape), even within individual cases (e.g., Figure 1). This visual complexity presents a key challenge but is crucial for training robust algorithms capable of discerning subtle differences for accurate classification.

In addition to image data, the ISIC 2024 dataset includes rich metadata describing patient demographics and lesion characteristics. Key fields cover diagnostic information such as lesion ID, melanoma thickness, mitotic index, age, sex, anatomical site, lesion size, and color or symmetry metrics. These features provide valuable clinical context that complements image-based information, enabling multimodal models to learn both visual and non-visual patterns. By integrating metadata with image features, models can achieve higher diagnostic accuracy, improved interpretability, and more reliable performance in real-world skin cancer detection.

2.2. CNNs for skin cancer detection

Convolutional neural networks (CNNs) have been widely adopted for skin lesion classification due to their ability to capture local features essential for distinguishing between benign and malignant lesions [12]. Early studies leveraged pre-trained CNN models such as ResNet [1], EfficientNet [2], and DenseNet [3], demonstrating notable success in skin cancer classification on datasets like ISIC. For example, in the SIIM-ISIC Melanoma Classification Challenge (2020), an ensemble model combining multiple CNN architectures, primarily based on EfficientNet, won the competition due to its high diagnostic accuracy in identifying melanoma [13]. These CNN-based methods laid the groundwork for the next generation of models that combine CNNs with other advanced architectures, thereby allowing for improved generalization and integration with additional data sources.

Previous research, such as the work by Ha et al. [13], applied an ensemble of CNN models to the ISIC 2020 dataset and achieved impressive diagnostic results by utilizing high-resolution images and primarily image-based features. However, when we transition to the ISIC 2024 dataset, significant limitations emerge. The images in ISIC 2024 are considerably smaller (under 256×256 pixels) compared to the larger images (around 1024×1024 pixels) used in previous datasets. This decrease in image resolution hampers the CNN's ability to capture fine local details, a strength of CNNs when working with high-resolution images [14]. Additionally, while efforts have been made to integrate patient metadata into CNN models, these methods typically use metadata in a limited manner. Despite the abundant metadata available in ISIC 2024 (55 features for training and 44 for testing), past studies have not fully utilized this rich source of diagnostic data [13].

CNNs are well-suited for extracting local features from high-resolution images but struggle to capture critical details when resolution decreases. In low-resolution settings such as ISIC 2024, metadata becomes an essential complementary source of diagnostic information. However, existing CNN-based methods have not fully leveraged this metadata or effectively addressed data imbalance, underscoring the need for multimodal approaches that integrate image and metadata features to improve diagnostic accuracy and robustness.

2.3. Transformers for skin cancer detection

Transformers, initially developed for natural language processing, have shown exceptional potential in visual tasks through self-attention mechanisms. The seminal work "Attention is

All You Need" [15] introduced the concept of self-attention, which enables models to learn long-range dependencies in data more effectively than traditional convolutional or recurrent layers. Vision Transformers (ViTs) extend this concept to the visual domain by dividing an image into patches and treating each patch as a token, allowing the model to capture global contextual information across the image [16]. Research indicates that Vision Transformers (ViTs) are effective in classifying medical images, particularly when utilizing methods such as data augmentation and synthetic image generation to tackle class imbalance. A main advantage of Transformers is that they are good at capturing long-range connections in data, helping to understand the overall context better, which is often hard for CNNs.

For instance, in skin cancer detection, Cai et al. [17] developed a multimodal Transformer that integrates image data and metadata using dual encoders, leading to improved accuracy in comparison to traditional CNNs. Another study, Krishna et al. [18], used ViT-GANs (Vision Transformers combined with Generative Adversarial Networks) to generate synthetic images for rare classes, thus balancing the dataset and improving model robustness. These studies demonstrate the efficacy of Transformer models in handling complex, imbalanced medical datasets, further supporting their use in multimodal approaches. However, careful consideration is needed regarding factors that may influence the performance of Transformers, especially ViTs, when applied to the ISIC 2024 dataset.

Transformer-based approaches, such as the multimodal model developed by Cai et al. [17] and the LesionAid study using ViT-GANs [18], have shown potential by combining image data with metadata and utilizing synthetic image generation to address data imbalance. Despite these promising results, when these approaches are applied to the ISIC 2024 dataset, significant limitations arise. Vision Transformers (ViTs) were originally designed for high-resolution image processing. Their performance, however, can be compromised when applied to the smaller images (under 256×256 pixels) common in ISIC 2024. Moreover, while some studies have incorporated metadata into Transformer models, this integration is often secondary to the focus on image-based features. This limited integration of metadata, which is especially rich in ISIC 2024, means that previous Transformer models may not fully leverage all the diagnostic information available.

Transformer models also encounter limitations in datasets like ISIC 2024. Although they effectively capture global dependencies through self-attention, dividing images into patches can lead to the loss of fine local details, especially in low-resolution images. Existing studies integrating metadata into Transformers [19] often treat it as auxiliary rather than essential, and approaches like ViT-GANs focus on image quality rather than data fusion. Hence, future research should develop Transformer-based frameworks that better integrate image and clinical metadata while addressing class imbalance for more accurate skin cancer classification.

2.4. Multimodal fusion for skin cancer detection

The integration of image data with metadata is increasingly recognized as a powerful approach for enhancing diagnostic performance in skin disease classification. Early studies primarily focused on image-only models, but recent research has demonstrated that combining metadata with image features can substantially improve classification accuracy [17]. In a recent work, Cai et al. [17] employed a multimodal Transformer architecture utilizing dual encoders for processing image and metadata inputs, along with a cross-attention mech-

anism in the decoder to fuse these heterogeneous features. This design achieved a notable improvement in diagnostic performance on the ISIC dataset by effectively leveraging both visual and clinical contextual information.

Similarly, Aladhadh et al. [20] introduced the Medical-ViT framework, which enriches skin lesion image representations with metadata to provide contextual insights. Their model achieved a test accuracy of 96.14% on the ISIC dataset, underscoring the critical importance of multimodal fusion in capturing complex correlations between visual and non-visual cues. Such fusion-based methods significantly enhance the robustness and diagnostic accuracy of automated systems for skin cancer detection, paving the way for next-generation computer-aided dermatological analysis.

2.5. Research gap

Despite notable advances in CNN, Transformer, and multimodal-based methods for skin cancer detection, several key challenges remain. Most existing models are optimized for high-resolution images, whereas datasets like ISIC 2024 contain predominantly low-resolution samples (below 256×256 pixels), which limit fine-grained feature extraction. Severe class imbalance (≈99.9% vs. 0.1%) further biases model learning, while current fusion strategies underutilize valuable clinical metadata by treating it as auxiliary information. To overcome these issues, this study introduces FusionNetX, a comprehensive multimodal framework designed to enhance feature representation, balance learning, and improve diagnostic reliability in skin cancer detection.

3. FUSIONNETX FRAMEWORK: A MULTIMODAL APPROACH FOR SKIN CANCER DETECTION

This study introduces FusionNetX (Figure 2), a robust multimodal framework that integrates image and metadata from the ISIC 2024 dataset. It consists of three core modules: (i) Data preprocessing and feature engineering, (ii) Multimodal feature extraction, and (iii) Ensemble classification. FusionNetX is specifically designed to address severe class imbalance and exploit complementary visual and clinical information to improve diagnostic accuracy.

3.1. Data preprocessing

Image data processing. Images from the ISIC 2024 dataset are resized to 224×224 pixels and normalized using ImageNet statistics to ensure input consistency and enable effective transfer learning. Data augmentation techniques, including random cropping, rotation, flipping, and color jittering, are applied to enhance generalization and mitigate overfitting by simulating real-world variations in lighting and orientation. These steps not only improve robustness and convergence stability but also optimize computational efficiency for advanced architectures such as ConvNeXtV2 and Vision Transformers within the FusionNetX framework.

Metadata processing. Metadata containing patient demographics and lesion characteristics undergoes preprocessing to enhance data quality and feature richness. Missing numerical values are imputed with the median, while categorical variables (e.g., gender, site, and lesion type) are encoded using One-Hot Encoding. Numerical features are normalized

per patient to reduce individual variability, and new features such as lesion size ratio and shape index are derived to capture clinical relevance. To address class imbalance, RandomOverSampler and RandomUnderSampler are employed, ensuring balanced representation and improving model robustness. Collectively, these preprocessing steps strengthen FusionNetX's ability to learn discriminative patterns and generalize effectively across diverse clinical scenarios.

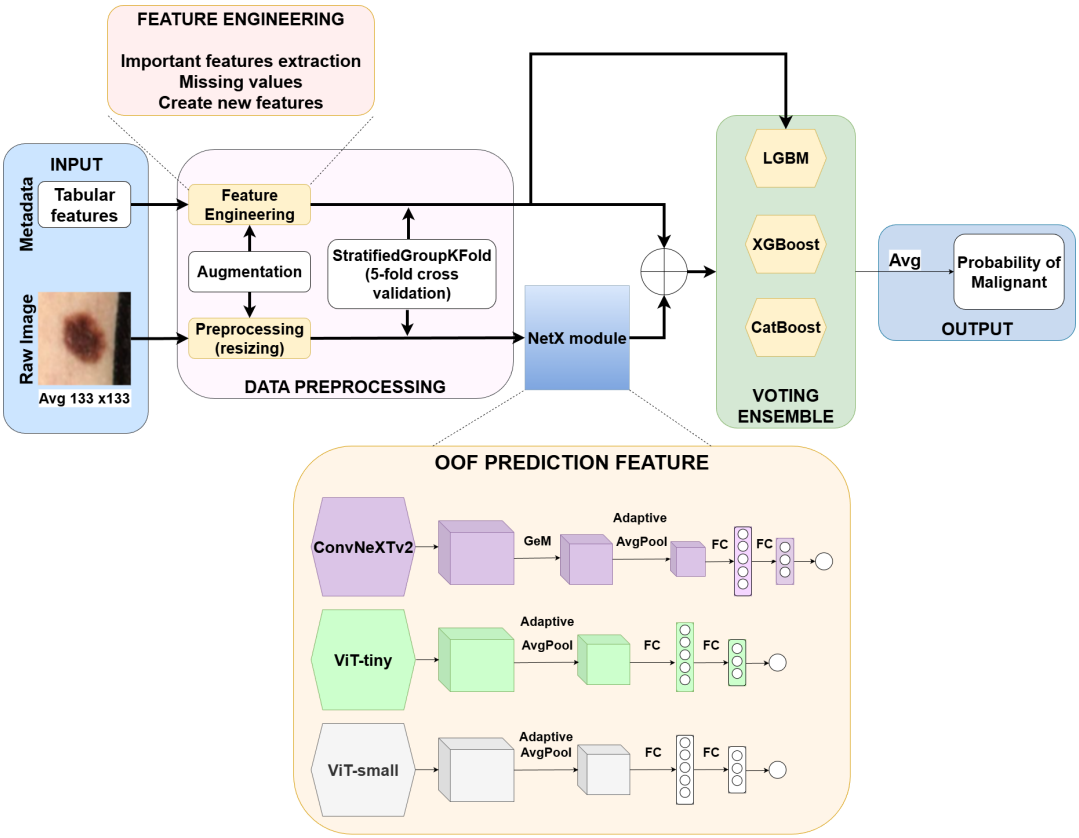

Figure 2: The FusionNetX framework for skin cancer detection

3.2. Out of folds prediction feature

The out-of-folds (OOF) prediction feature module serves as the cornerstone of the FusionNetX model, meticulously engineered to extract rich, meaningful representations from image data. This module comprises two key components: Image Feature Extraction via CNNs and Transformers, and GeM and Adaptive Average Pooling. Together, these components enhance the model's ability to capture both fine-grained and broad patterns in skin lesion images, which are crucial for accurate classification. The combination of advanced architectures like ConvNeXtV2 and Vision Transformer (ViT) allows for comprehensive feature extraction. At the same time, GeM Pooling offers flexible, dynamic aggregation of features, and Adaptive Average Pooling standardizes output sizes to ensure consistency. This synergy improves the model's robustness, reduces overfitting, and enables it to generalize well to unseen data, ultimately ensuring high accuracy in skin cancer detection tasks.

3.2.1. Image feature extraction via CNNs and transformers

At the image feature extraction stage, preprocessed lesion images are fed into multiple state-of-the-art architectures, including ConvNeXtV2 [21] and vision transformers (ViT-tiny and ViT-small) [16]. ConvNeXtV2, a convolutional network inspired by Transformer design, captures local spatial patterns, while ViT models learn global contextual relationships through attention mechanisms. All models are fine-tuned on lesion images, and their outputs are passed through fully connected layers to distill high-level features for subsequent classification.

Multi-scale pooling is applied in ConvNeXtV2 to aggregate features captured at different receptive fields using GeM Pooling [22]. By combining features from multiple kernel sizes (e.g., 31×31 and 3×3) [23], the model effectively learns both fine-grained details and broader structural patterns. The aggregated feature map is then processed through an Adaptive Average Pooling layer to standardize the output size (typically 1×1), ensuring consistency regardless of input resolution. This combination enables ConvNeXtV2 to capture multi-scale representations, enhancing its ability to detect lesions of varying sizes and shapes.

$$X_{\text{multi-scale}} = \sum_{k\in\{3,5\}} w_k \cdot \text{GeM}_k(X), \tag{1}$$

where w_k represents learnable weights that are specifically assigned to each kernel size k, allowing the model to adaptively weigh the contribution of features from different scales. $\text{GeM}_k(X)$ denotes the output obtained by applying Generalized Mean (GeM) pooling with a kernel size k to the input feature map X. The output from the multi-scale pooling stage, $X_{\text{multi-scale}}$, then undergoes adaptive average pooling, formulated as

$$X_{\text{pooled}} = \text{AdaptiveAvgPool2d}\big(X_{\text{multi-scale}}, \text{output\_size} = (1,1)\big). \tag{2}$$

Here, X_{pooled} represents the final output feature map after adaptive pooling, which is standardized to a specified output size, in this case $(1, 1)$, effectively creating a global average representation of the multi-scale features.

ConvNeXtV2 employs GeM and Adaptive Average Pooling to capture subtle structural variations in lesions, while ViT-tiny and ViT-small utilize self-attention to model global contextual relationships across image regions. Together, these architectures complement each other, ConvNeXtV2 excels at extracting fine-grained local details through convolution and multi-scale pooling, whereas ViT captures long-range dependencies essential for understanding lesion morphology.

Both networks incorporate normalization, residual connections, and adaptive pooling for stable and robust training on diverse medical images. After feature extraction, outputs from each model are passed through fully connected layers to generate out-of-fold predictions during StratifiedGroupKFold cross-validation, which are subsequently used as inputs for the final ensemble classifier.

3.2.2. GeM and adaptive average pooling

In deep learning models for skin lesion classification, pooling layers play a crucial role in capturing discriminative features across multiple scales. FusionNetX employs GeM and

Adaptive Average Pooling to enhance feature extraction from dermoscopic images, emphasizing subtle details vital for distinguishing malignant from benign lesions.

GeM [23] generalizes traditional pooling methods such as average and max pooling by introducing a learnable parameter p that controls the degree of aggregation during pooling

$$\mathbf{e} = \left[\left(\frac{1}{|\Omega|} \sum_{u \in \Omega} x_{c,u}^{p} \right)^{\frac{1}{p}} \right]_{c=1,\ldots,C}. \tag{3}$$

GeM pooling allows FusionNetX to adaptively control the level of detail captured from lesion images, emphasizing salient textures and regions critical for malignancy detection. It balances fine-grained feature retention with broader contextual understanding, consistent with prior findings on GeM's effectiveness in image classification and retrieval [22]. Complementarily, Adaptive Average Pooling standardizes feature map dimensions regardless of input resolution, mitigating information loss found in standard pooling and preserving fine details for precise classification [24]. Together, these pooling layers ensure stable training and produce consistent, detail-rich feature representations, which are essential for robust medical image analysis.

3.3. Features fusion

After image features are extracted from multiple image models, these features are then combined with metadata to create a comprehensive representation for skin cancer classification. Specifically, after running the image models through Out-of-fold cross-validation, each image model will produce predictions across five folds. In total, there are four different image models applied to the dataset, and each model's predictions will be organized into five folds, resulting in four columns of features. These four columns represent the predicted probabilities of the image models (in terms of the likelihood of malignancy) for each of the five folds.

Mathematically, for each image model i, the prediction in fold j for an image x_k (where k indexes each image) can be represented as $P_{ij}(x_k)$, where $P_{ij}(x_k) \in [0, 1]$ is the probability of malignancy predicted by the i-th image model for the j-th fold of image x_k.

Given four image models, the final set of image feature vectors for an image x_k across five folds will be

$$F_{\text{image}}(x_k) = [P_{1j}(x_k),\ P_{2j}(x_k),\ P_{3j}(x_k),\ P_{4j}(x_k)] \text{ for } j = 1, 2, 3, 4, 5. \tag{4}$$

This results in four columns of feature values for each image x_k, corresponding to the predictions across the five folds of the four image models. These sets of features $F_{\text{image}}(x_k)$ are then concatenated with the metadata features $M(x_k)$, which include clinical information such as age, gender, and other patient-specific attributes. The complete feature vector for an image x_k is then formed as

$$F_{\text{fusion}}(x_k) = [F_{\text{image}}(x_k),\ M(x_k)]. \tag{5}$$

The final fusion of features $F_{\text{fusion}}(x_k)$ combines both image features (predicted probabilities of malignancy) and metadata, ready for further processing (e.g., classification or regression) by a machine learning model.

This Multimodal Fusion approach enables the model to leverage the complementary information from both image and metadata, to improve classification accuracy. By combining predictions from multiple image models with rich clinical metadata, the model can compensate for the weaknesses of individual sources of data, especially in cases where image resolution is low, as is common with the ISIC 2024 dataset. This fusion strategy allows the model to focus not only on the image features but also on the underlying clinical context, improving its robustness in the diagnostic process.

3.4. Voting ensemble

In this final stage, both the image-based OOF predictions and the metadata features are combined through a Voting Ensemble. The formula for the ensemble prediction P_e is

$$P_e = \frac{1}{N} \sum_{i=1}^{N} P_{\text{model}_i}, \tag{6}$$

where P_e is the final ensemble prediction probability, P_{model_i} is the prediction probability from each model (e.g., LightGBM, XGBoost, and CatBoost, and N is the number of models in the ensemble (in this case, $N = 3$).

The Voting Ensemble in FusionNetX mitigates individual model biases by averaging predictions across multiple models and cross-validation folds. This dual-averaging strategy enhances robustness, minimizes overfitting, and improves generalization to unseen data. By leveraging the collective intelligence of diverse learners, FusionNetX achieves higher reliability and predictive accuracy, making it well-suited for real-world skin cancer detection applications.

4. EXPERIMENT AND RESULTS

4.1. Experiment setup

Dataset and task. Experiments were conducted on the ISIC 2024 Kaggle dataset containing 401,059 dermatoscopic images with an extreme benign-to-malignant imbalance (≈1000:1). The task is binary classification, distinguishing benign from malignant lesions, using both images and metadata. FusionNetX combines CNN-based image features with gradient boosting decision trees (GBDT) to improve tabular learning under imbalance. Model performance was assessed via 5-fold cross-validation (CV) following top ISIC 2024 competition practices. As the official test labels remain undisclosed, evaluation relied on pAUC scores submitted to the Kaggle platform.

Experimental environment. All experiments were executed on Kaggle using NVIDIA P100 GPUs. The implementation employed *PyTorch*, *scikit-learn*, and *Optuna* [25], ensuring a consistent and reproducible setup.

Data partitioning and validation. A 5-fold `StratifiedGroupKFold` strategy was adopted to prevent data leakage, maintaining class balance while grouping all images of each patient ($patient_{id}$) within the same fold. Training and validation used only the provided training set, with the public test set (28%) used for feedback and the private test set (72%) for final leaderboard evaluation.

Model design and training. To extract key visual features from dermatoscopic images, three architectures were utilized: ConvNeXtV2-Nano, ViT-Tiny, and ViT-Small, combining convolutional neural networks (CNNs) and vision transformers (ViTs). These models were selected for their proven efficiency and effectiveness in image recognition tasks, balancing model complexity with computational feasibility for large datasets like ISIC 2024. The training hyperparameters for these models are detailed in Table 1.

Table 1: Hyperparameters for image models

Parameter	ConvNeXtV2	ViT-Tiny	ViT-Small
Model name	convnextv2_nano	vit_tiny_patch16	vit_small_patch16
Learning rate (LR)	1e-4	1e-4	1e-4
LR decay	–	1.0	1.0
Batch size	32	32	32
Image size	224	224	224
Epochs	30	30	30
Scheduler LR	CosineAnnealingLR	CosineAnnealingLR	CosineAnnealingLR
T_max (Scheduler)	500	500	500
Warmup Ratio	0.05	0.05	0.05
Min LR	1e-6	1e-6	1e-6

To address specific dataset nuances, two specialized ViT models (Tiny and Small) were trained on a curated subset $\mathcal{C}$, defined as

$$\mathcal{C} = \{x \mid y = 1,\ \text{iddx}_1 = \text{label},\ \text{iddx}_2 \neq \text{NaN}\},\ \text{followed by } y = \text{target}. \tag{7}$$

This subset targets malignant cases with diagnostic metadata (iddx_1 and iddx_2), prevalent in the training set but potentially underrepresented in the test set, thus enhancing model robustness for these patterns.

Table 2: Hyperparameters for tree-based models

Parameter	LightGBM	CatBoost	XGBoost
Objective	binary	Logloss	binary
Boosting Type	gbdt	–	hist
Random State	42	42	42
Learning Rate	0.03231	0.06936	0.08501
Max Depth	4	7	6
Subsample	–	0.6249	0.6013
Number of Leaves	103	–	–
Colsample by Tree	0.83296	–	0.84378
Min Data in Leaf	85	24	85
Scale Pos Weight	2.7984	2.6149	3.2944
Device	cpu	gpu	cpu
Iterations	200	250	100

FusionNetX integrates high-performance tree-based classifiers, LightGBM, CatBoost, and XGBoost, from `scikit-learn` for metadata processing. These models were chosen for their efficiency on tabular data, robustness to noise, and ability to model non-linear feature interactions. Hyperparameters were optimized using Optuna (Table 2). To mitigate class imbalance, oversampling and undersampling were applied alongside comprehensive feature engineering, including missing-value imputation, categorical encoding, and interaction feature creation.

FusionNetX employs an ensemble of four models: (i) LightGBM (metadata-only), (ii) LightGBM (image-enhanced metadata), (iii) CatBoost (image-enhanced metadata), and

(iv) XGBoost (image-enhanced metadata). Final predictions are aggregated using a VotingClassifier with soft voting, which improves accuracy and stability on imbalanced data. This multimodal ensemble design enhances diagnostic performance and generalization on the ISIC 2024 dataset.

4.2. Evaluation metrics

The primary evaluation metric for this study is the *partial area under the Receiver Operating Characteristic curve (pAUC)*, which measures model performance within a clinically significant range of the ROC curve. The standard area under the curve (AUC) quantifies a classifier's overall ability to distinguish between positive and negative cases across all thresholds, with higher values indicating stronger discrimination. In contrast, pAUC focuses on a specific high-sensitivity region, typically where the true positive rate (TPR) exceeds 80%, to emphasize clinical relevance. This is particularly important in cancer detection, where minimizing false negatives is critical. A larger pAUC value therefore indicates a model that maintains high sensitivity while effectively controlling false positives.

To provide a more comprehensive evaluation, we also report Accuracy and Average Precision (AP). Accuracy measures the overall proportion of correctly classified samples and is calculated as

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}.$$

Average Precision (AP) summarizes the precision, recall relationship by averaging precision values at different recall levels, where

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

Mathematically, AP is expressed as

$$AP = \sum_{n} (R_n - R_{n-1}) P_n, \tag{8}$$

where P_n denotes precision at recall level R_n. AP provides a more informative assessment for imbalanced datasets, capturing the trade-off between precision and recall across multiple thresholds.

4.3. Impact of advanced processing metadata techniques

To highlight the effectiveness of advanced metadata handling, we compared two preprocessing approaches in Table 3 including a basic method using simple target encoding and missing value imputation, which achieved pAUC scores of 0.15170 (public) and 0.14290 (private), and the sophisticated techniques in FusionNetX, which incorporated enhanced missing value handling and advanced feature engineering, significantly improving pAUC scores to 0.18024 (public) and 0.16040 (private).

This significant performance difference underscores the value of advanced metadata processing in effectively leveraging non-image information to enhance model accuracy and robustness, particularly in scenarios where visual data may be insufficient.

Table 3: Comparison of performance between simple and advanced metadata preprocessing methods.

Method	Public Score	Private Score
Simple metadata processing	0.15170	0.14290
Advanced metadata processing (FusionNetX)	0.18024	0.16040

4.4. Evaluating image and metadata contributions

To understand the contribution of each component in our framework, we conducted ablation studies by selectively removing image features or metadata features. As shown in Table 4, the results confirm that both modalities contribute substantially to overall performance. Image features provide the primary discriminative power, while metadata features complement them by boosting classification accuracy and enhancing the model's ability to make accurate predictions.

Table 4: The results of an ablation study

Model Variant	Public Score	Private Score
Full multimodal model	0.18380	0.17295
Image-only model	0.15525	0.14151
Metadata-only model	0.18024	0.16040

Image features. The image-only model achieved pAUC scores of 0.15525 (public) and 0.14151 (private), showing that image features like pixel values, texture, and color patterns are key for distinguishing malignant from benign lesions. However, its performance was lower than the full multimodal model, indicating limitations in cases with visually ambiguous or indistinct lesion characteristics.

Metadata features. In contrast, the metadata-only model utilizes non-image features like patient age, gender, and lesion location. While these results are lower than the full multimodal model, they demonstrate that metadata provides valuable supplementary information, improving performance compared to the image-only model. Metadata enables the model to learn contextual patterns and correlations not directly visible in the images, highlighting the importance of clinical context. However, the results also indicate that metadata alone is not sufficient to reach the performance level of the full multimodal model, emphasizing the benefit of combining it with image data.

Synergistic impact of combining. The full multimodal model significantly outperformed both the image-only and metadata-only models. This demonstrates the synergistic effect of combining image and metadata features. Image features provide detailed visual information crucial for lesion classification, while metadata offers contextual knowledge that enables more informed predictions. This combination allows the model to handle a broader range of cases, including those where visual data alone might be insufficient.

The ablation studies highlight the importance of integrating both image and metadata features. The full multimodal model outperformed both image-only and metadata-only models, demonstrating the synergistic effect of combining these data sources. Additionally, removing the ensemble voting mechanism resulted in a noticeable drop in performance, emphasizing the value of ensemble learning in enhancing model robustness.

4.5. Experimental results on ISIC dataset

Table 5 summarizes the pAUC scores for the top 10 teams in the ISIC 2024 Challenge [7], including our proposed framework, FusionNetX. The rankings are based on private scores, which are calculated on a hidden test dataset to ensure fair evaluation and assess model performance on unseen data. The public scores, derived from a separate validation set, provide additional insight into the models' capabilities during the competition. This table highlights the competitive nature of the challenge and the high performance of the participating teams.

The results in Table 5 reveal that FusionNetX, our proposed framework, achieved the highest private score (0.17295), demonstrating superior generalization performance compared to the other top teams. Despite having a slightly lower public score (0.18380) than some competitors, FusionNetX outperformed them on the private leaderboard, which is the primary metric for ranking. The second-place team, Ilya Novoselskiy, followed closely with a private score of 0.17273, highlighting a marginal difference. Notably, Sinan Calisir, who achieved the highest public score (0.18734), ranked 10th in the private score, emphasizing the importance of evaluating models on unseen data to avoid overfitting.

Table 5: pAUC scores sorted by private score for top 10 teams in the ISIC 2024 challenge, including the proposed framework.

No	Team	Public Score	Private Score
1	FusionNetX (our model)	0.18380	0.17295
2	Ilya Novoselskiy	0.18611	0.17273
3	Yakiniku	0.18635	0.17243
4	KS	0.18421	0.17229
5	BiBanhBao	0.18274	0.17225
6	Kanna Hashimoto friends 2	0.18323	0.17210
7	xck	0.18259	0.17192
8	Former ZLP-DSs	0.18498	0.17163
9	Ujjwal Pandey	0.18078	0.17158
10	Sinan Calisir	0.18734	0.17146

Many of the top teams in the ISIC 2024 Challenge adopted similar methodologies, centering on multimodal fusion of image features and patient metadata, combined with ensemble learning and hyperparameter tuning. For example, the runner-up team led by Ilya Novoselskiy focused heavily on synthetic data augmentation to address class imbalance but placed less emphasis on fine-tuning their models, which may have limited their private score (0.17273). Another strong competitor, the Yakiniku team, employed a large-scale ensemble of 54 tree-based models with extensive feature engineering. While effective in many respects, such complexity may have introduced redundancy and reduced generalization to unseen data.

The results highlight that our proposed FusionNetX framework effectively leverages the combined strengths of image and metadata features to address the challenges of extreme class imbalance in the ISIC 2024 dataset. By integrating deep learning-based image feature extraction with tree-based ensemble models like LightGBM, CatBoost, and XGBoost, FusionNetX surpassed approaches relying solely on image or metadata data. This multimodal design proved crucial for achieving superior performance, particularly in clinically significant high-sensitivity regions. The consistent alignment between public and private scores further validates the framework's robust generalization capabilities, positioning it as a reliable solution for skin cancer detection.

4.6. Experimental results on HAM10000 and PH2 datasets

To assess the generalizability of FusionNetX, additional experiments were conducted on two benchmark datasets: HAM10000 and PH2. These evaluations tested the model's robustness across different lesion types and real-world conditions. The experimental setup followed the same methodology as in ISIC 2024, with minor hyperparameter adjustments: 20 training epochs and a batch size of 128 for HAM10000, and 80 epochs with a batch size of 64 for PH2. These settings ensured stable training and optimal performance on datasets of varying size and complexity.

Table 6: Comparison results on HAM10000 dataset

Model	Accuracy	AUC	Average Precision
IRv2+soft attention	93.4	0.984	0.937
FusionNetX (our model)	95.2	0.995	0.885

The FusionNetX model achieves superior performance on the HAM10000 dataset, surpassing the IRv2+Soft Attention model [26] in terms of both accuracy and AUC. As shown in Table 6, while the average precision of FusionNetX is slightly lower than IRv2+Soft Attention, its significantly higher AUC and accuracy suggest a more robust performance in terms of overall classification accuracy and the model's ability to distinguish malignant from benign lesions.

Table 7: Results on PH2 dataset

Model	Accuracy	AUC	Average Precision
FusionNetX (our model)	95.0	0.9844	0.9663

Similarly, on the PH2 dataset, as shown in Table 7, FusionNetX demonstrates high accuracy, achieving an AUC of 0.9844 and an average precision of 0.9663. These results highlight the model's strong ability to generalize across different datasets, further validating its robustness and effectiveness in skin lesion classification tasks.

The ablation study highlights the complementary roles of image and metadata features in FusionNetX. Image features provide the necessary discriminative power, while metadata enhances accuracy and robustness, especially in challenging cases. The full multimodal model, combining both features, achieves the best performance, demonstrating their synergistic impact. The ensemble learning mechanism further improves robustness.

Experiments on the HAM10000 and PH2 datasets confirm the generalizability of FusionNetX, validating its effectiveness across different skin lesion classification tasks. These results address the reviewers' concerns and underscore the model's potential for real-world clinical applications in dermatology.

5. CONCLUSION AND FUTURE WORK

This study developed FusionNetX, a multimodal framework integrating deep learning-based image feature extraction with tree-based ensemble classifiers for skin cancer detection using the ISIC 2024 dataset. Our framework effectively addresses the challenges of extreme class imbalance and ensures model generalization through sophisticated cross-validation techniques. The achieved pAUC scores on both cross-validation and private test sets demonstrate

the potential of our model for practical deployment in clinical settings, aiding in the early and accurate detection of malignant skin lesions. Notably, FusionNetX secured the top position on the private leaderboard, achieving the highest private score (0.17295) among all competitors. This top-ranking performance underscores the framework's ability to generalize effectively to unseen data, further highlighting its robustness and reliability for real-world clinical applications. Furthermore, testing on external datasets, including HAM10000 and PH2, confirmed its robustness across various skin lesion types. These outstanding results reinforce FusionNetX's role in advancing skin cancer diagnosis, ultimately improving patient prognosis and treatment opportunities.

Although FusionNetX achieved strong performance, its reliance on the ISIC 2024 dataset limits the generalizability of the results. The absence of a public test set and restricted computational resources on Kaggle hindered the evaluation of larger models and broader comparisons. Moreover, the competition's emphasis on pAUC excluded key clinical metrics such as sensitivity and accuracy. These constraints affected comprehensive validation, but the findings still demonstrate the model's robustness and potential for clinical deployment.

Future research will address these limitations by employing synthetic data generation to improve robustness and exploring advanced models such as MedViT and BEiT for enhanced feature extraction. Additional evaluations using sensitivity and accuracy will ensure better clinical relevance. We also plan to refine ensemble strategies, notably, a seven-model LightGBM ensemble already improved performance (pAUC 0.18131 public, 0.16581 private), suggesting further potential for optimization and real-world application.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778, https://doi.org/10.1109/CVPR.2016.90.

[2] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. on Machine Learning*. PMLR, 2019, pp. 6105–6114, https://proceedings.mlr.press/v97/tan19a/tan19a.pdf.

[3] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708, https://doi.org/10.1109/CVPR.2017.243.

[4] M.-T. Vo, A.-H. Vo, and T. Le, "A robust framework for shoulder implant x-ray image classification," *Data Technologies and Applications*, vol. 56, no. 3, pp. 447–460, 2022, https://doi.org/10.1108/DTA-08-2021-0210.

[5] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, no. 1, pp. 1–9, 2018, https://doi.org/10.1038/sdata.2018.161.

[6] T. Mendonça, M. E. Celebi, T. Mendonça, and J. Marques, "PH2: A public database for the analysis of dermoscopic images," in *Dermoscopy Image Analysis*, 2015, pp. 91–111, https://repositorio-aberto.up.pt/bitstream/10216/110404/2/215668.pdf.

[7] International Skin Imaging Collaboration, "SLICE-3D 2024 challenge dataset," https://doi.org/10.34970/2024-slice-3d, 2024.

[8] N. Kurtansky, V. Rotemberg, M. Gillis, K. Kose, W. Reade, and A. Chow, "ISIC 2024 - skin cancer detection with 3D-TBP," Kaggle, 2024, https://kaggle.com/competitions/isic-2024-challenge.

[9] N. R. Kurtansky, B. M. D'Alessandro, M. C. Gillis, and et al., "The SLICE-3D dataset: 400,000 skin lesion image crops extracted from 3D TBP for skin cancer detection," *Scientific Data*, vol. 11, p. 884, 2024, https://doi.org/10.1038/s41597-024-03743-w.

[10] C. Horsham, M. Janda, M. Kerr, H. P. Soyer, and L. J. Caffery, "Consumer perceptions on privacy and confidentiality in dermatology for 3D total-body imaging," *Australasian Journal of Dermatology*, vol. 64, pp. 118–121, 2023, https://doi.org/10.1111/ajd.13952.

[11] A. M. Ferry, S. M. Sarrami, P. C. Hollier, C. F. Gerich, and J. F. Thornton, "Treatment of non-melanoma skin cancers in the absence of mohs micrographic surgery," *Plastic and Reconstructive Surgery – Global Open*, vol. 8, p. e3300, 2022, https://doi.org/10.1097/GOX.0000000000003300.

[12] R. Archana and P. S. E. Jeevaraj, "Deep learning models for digital image processing: A review," *Artificial Intelligence Review*, vol. 57, p. 11, 2024, https://doi.org/10.1007/s10462-023-10631-z.

[13] Q. Ha, B. Liu, and F. Liu, "Identifying melanoma images using efficientnet ensemble: Winning solution to the SIIM-ISIC melanoma classification challenge," *arXiv preprint*, 2020, https://arxiv.org/abs/2010.05351.

[14] O. Rukundo, "Effects of image size on deep learning," *Electronics*, vol. 12, p. 985, 2023, https://doi.org/10.3390/electronics12040985.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, https://dl.acm.org/doi/10.5555/3295222.3295349.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021, https://arxiv.org/abs/2010.11929.

[17] G. Cai, Y. Zhu, Y. Wu, X. Jiang, J. Ye, and D. Yang, "A multimodal transformer to fuse images and metadata for skin disease classification," *Visual Computer*, vol. 39, no. 7, pp. 2781–2793, 2023, https://doi.org/10.1007/s00371-022-02492-4.

[18] G. S. Krishna, K. Supriya, and M. Sorgile, "LesionAid: Vision transformers-based skin lesion generation and classification," *arXiv preprint*, 2023, https://arxiv.org/abs/2302.01104.

[19] S. Vachmanus, T. Noraset, W. Piyanonpong, T. Rattananukrom, and S. Tuarob, "DeepMetaForge: A deep vision-transformer metadata-fusion network for automatic skin lesion classification," *IEEE Access*, vol. 11, pp. 145 467–145 484, 2023, https://doi.org/10.1109/ACCESS.2023.3345225.

[20] S. Aladhadh, M. Alsanea, M. Aloraini, T. Khan, S. Habib, and M. Islam, "An effective skin cancer classification mechanism via medical vision transformer," *Sensors*, vol. 22, no. 11, p. 4008, 2022, https://doi.org/10.3390/s22114008.

[21] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, "Convnext v2: Co-designing and scaling convnets with masked autoencoders," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 16 133–16 142, https://doi.org/10.48550/arXiv.2301.00808.

[22] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning CNN image retrieval with no human annotation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1655–1668, 2018, https://arxiv.org/abs/1711.02512.

[23] X. Ding, X. Zhang, J. Han, and G. Ding, "Scaling up your kernels to 31×31: Revisiting large kernel design in CNNs," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11 963–11 975, https://arxiv.org/abs/2203.06717.

[24] J. Yang, F. Chen, R. K. Das, Z. Zhu, and S. Zhang, "Adaptive-avg-pooling based attention vision transformer for face anti-spoofing," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2024, pp. 3875–3879, https://arxiv.org/abs/2401.04953.

[25] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2019, pp. 2623–2631, https://doi.org/10.1145/3292500.3330701.

[26] S. K. Datta, M. A. Shaikh, S. N. Srihari, and M. Gao, "Soft attention improves skin cancer classification performance," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2021, pp. 13–23, https://doi.org/10.1007/978-3-030-87444-5_2.

Received on November 28, 2024
Accepted on September 29, 2025