Int. J. Nonlinear Anal. Appl. In Press, 1–17
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2025.37710.5468

Meta-Classifier Fusion of Deep Ensemble Representations for Enhanced Skin Lesion Diagnosis

S. Talayeh Tabibi[a], Alireza Nikravanshalmani[a,*], Hadi Saboohi[a]

[a]*Department of Computer Engineering, Ka.C, Islamic Azad University, Karaj, Iran*

(Communicated by Ehsan Kozegar)

Abstract

Although ensemble learning has proven to be effective for skin lesion classification, existing methods often underuse the complementary information embedded in heterogeneous convolutional neural networks (CNNs). In this work, we propose a meta-classifier framework that leverages high-dimensional feature representations from a frozen, diversified ensemble of CNNs to improve diagnostic robustness. To exploit both decision-level and feature-level diversity, we extract penultimate (pre-output) layer embeddings from each ensemble member, concatenate them into a unified feature vector, and train a meta-classifier on this enriched representation. We evaluated multiple meta-classifier architectures, including lightweight neural networks and kernel-based models, and benchmarked their performance against conventional ensemble averaging strategies. Experiments on the ISIC 2018 dataset (13,000+ dermoscopic images) demonstrate that our approach outperforms both individual CNNs and classical ensemble methods, achieving an accuracy of 91.32% (versus 90.15% in prior work). The meta-classifier's ability to learn cross-model feature interactions and suppress ensemble uncertainty is validated through ablation studies and robustness tests against noisy input. This work highlights the untapped potential of hierarchical fusion in ensemble-based medical image analysis.

Keywords:
2020 MSC: Ensemble Learning; Meta-Classifier; Dermoscopic Image Analysis; Feature Fusion; Skin Lesion Classification

*Corresponding author
Email addresses: s.talayeh.tabibi@gmail.com (S. Talayeh Tabibi), nikravan@kiau.ac.ir (Alireza Nikravanshalmani), saboohi@kiau.ac.ir (Hadi Saboohi)

Received: May 2025 *Accepted:* July 2025

1 Introduction

Skin cancer remains a global health crisis, with melanoma alone responsible for over 60,000 annual deaths worldwide [37]. The World Health Organization estimates that 2–3 million non-melanoma and 132,000 melanoma skin cancer cases emerge yearly, with rising incidence rates linked to environmental and demographic factors [22]. In the U.S., melanoma ranks as the fifth most common cancer, disproportionately affecting adults aged 25–29 [5], while countries like Australia face incidence rates quadruple those of Canada or the UK [16]. Early diagnosis is critical: localized melanoma has a 99% 5-year survival rate, which plummets to 30% once metastasized [30, 4]. However, visual screening by dermatologists remains subjective and expertise-dependent [8], driving demand for automated diagnostic systems to reduce human error and variability.

Traditional computer-aided diagnosis (CADx) systems relied on handcrafted feature extraction (e.g., texture, asymmetry) followed by machine learning classifiers like SVMs [36, 15]. While effective for binary classification (benign vs. malignant), these methods struggled with multi-class differentiation due to inter-class similarities and intra-class heterogeneity [14]. Deep convolutional neural networks (CNNs) revolutionized the field by automating feature learning, achieving state-of-the-art results in dermoscopic image analysis [19, 45]. Yet, single-model CNNs remain constrained by dataset bias and limited generalizability, prompting the adoption of ensemble methods to aggregate predictions across diverse architectures [22].

Recent ensemble strategies for skin lesion classification have prioritized diversity through data augmentation, architectural heterogeneity, or statistical disagreement metrics like Cohen's Kappa [35, 34]. While such approaches improve robustness, they primarily operate at the decision level (e.g., majority voting), discarding richer feature-level representations from intermediate CNN layers. This underutilizes the spatial and semantic patterns learned by individual models, which could synergistically enhance diagnostic precision. For instance, one CNN might excel at capturing global lesion morphology, while another detects local texture irregularities—features that, when fused, could resolve ambiguous cases.

In this work, we propose a hierarchical meta-classifier framework that leverages feature embeddings from a diversified CNN ensemble to surpass the limitations of decision-level aggregation. Unlike conventional ensembles, we extract high-dimensional feature vectors from the penultimate layers of frozen, architecturally heterogeneous CNNs, concatenate them into a unified representation, and train a meta-classifier to learn cross-model feature interactions. We evaluate multiple meta-classifier designs (e.g., lightweight DNNs, kernel-based models) on the ISIC 2018 dataset, demonstrating that feature fusion outperforms traditional voting mechanisms. Our key contributions include:

- A systematic methodology to maximize ensemble diversity at both data (bootstrapping) and feature (architectural heterogeneity) levels.
- A meta-classifier architecture that learns from complementary feature embeddings rather than aggregated predictions.

- Empirical validation showing improved robustness to noisy inputs and inter-class ambiguity, achieving 91.32% accuracy—a 1.17% gain over prior decision-level ensembles.

In order to meet the study's goals, this paper is organized as follows: Section 2 reviews the existing literature on skin lesion classification, focusing on ensemble learning and feature fusion strategies in deep learning. Section 3 details the ISIC 2018 dataset, preprocessing steps, and the proposed meta-classifier framework, including architectural details of the base CNNs and feature concatenation methodology. Section 4 presents quantitative performance metrics, comparing the meta-classifier against traditional decision-level ensembles and state-of-the-art methods. Section 5 analyzes the effectiveness of feature-level fusion, contrasts it with conventional voting strategies, and evaluates computational tradeoffs. Finally, Section 6 summarizes the advancements achieved through hierarchical feature fusion and outlines future directions for improving robustness and interpretability in automated skin cancer diagnosis.

2 Related Works

Ensemble methods for skin lesion classification have been extensively studied. Nyiri and Kiss [31] demonstrated that combining models with varied hyperparameters and preprocessing improves performance on ISIC datasets. Harangi [19] fused outputs from heterogeneous CNNs via weighted averaging, highlighting the benefits of architectural diversity. Recent work by Baykal Kablan and Ayas [7] combines ConvNeXt [25] models (Tiny, Small, Base, Large) trained on ISIC 2019. They fuse confidence scores from each model, achieving 97.7% accuracy. This highlights the power of scaling model size diversity within the same architecture family. Researchers in [29] proposed a Max Voting (MV) ensemble (Random Forest, MLPN, SVM) trained on the selected characteristics of the Genetic Algorithm (GA) from HAM10000 and ISIC 2018 achieved 94.7% precision. The GA balances feature relevance and model diversity, addressing class imbalance.

While general ensemble methods have demonstrated significant improvements in skin lesion classification, their reliance on static aggregation strategies (e.g., averaging, voting) limits their ability to fully exploit the complementary information embedded in heterogeneous base models. Recent advances in meta-classifier design offer a promising alternative by learning adaptive fusion mechanisms that capture cross-model interactions and suppress ensemble uncertainty. Unlike conventional approaches that operate solely on decision-level outputs (e.g., confidence scores or class probabilities), meta-classifiers leverage high-dimensional feature representations from base models, enabling more nuanced and context-aware fusion. This shift from static to learned aggregation has the potential to unlock new levels of diagnostic robustness, particularly in complex medical imaging tasks where feature-level diversity is critical. Notably, Goyal and Rajapakse [18] introduced a meta-learning layer using 1×1 convolution to fuse features from Inception-v4 and InceptionResnet-v2, pioneering learned fusion strategies for ensembles. Xie et al. [43] extended this with adaptive weighting in their multi-level ensemble, while Serte and Demirel [22] fused decisions from directional sub-band CNNs using summation rules. Gessert et al. [17] emphasized meta-learning for final predictions in class-imbalanced settings,

combining DenseNet, SENet, and ResNeXt outputs. Other works explore diverse architectures (e.g., EfficientNet [4], vision transformers [9]) and multi-modal fusion [32], though few address advanced meta-classifier design. While Kuncheva [23] theoretically established the necessity of diversity in ensembles, most existing methods rely on simple fusion (e.g., averaging, weighted sums) rather than trainable meta-classifiers.

While confidence-score fusion methods, such as those employed in [7], have demonstrated effectiveness by aggregating decision-level outputs from diverse base models, they inherently limit the exploitation of complementary information embedded in intermediate feature representations. These methods rely on static aggregation rules (e.g., averaging or weighted sums), which cannot adapt to the unique strengths and weaknesses of individual models or suppress uncertain predictions effectively. In contrast, our meta-classifier framework leverages high-dimensional feature-level embeddings from the penultimate layers of base models, enabling adaptive and context-aware fusion. By concatenating these embeddings and training a meta-classifier, our approach captures cross-model interactions and prioritizes reliable features, resulting in improved diagnostic robustness and generalization. Furthermore, while confidence-score fusion offers computational simplicity, it lacks interpretability and fails to provide insights into how decisions are aggregated. Our framework, supported by ablation studies and robustness tests, not only achieves superior accuracy but also enhances interpretability by revealing the contributions of feature-level diversity to the final decision. This hierarchical fusion approach underscores the untapped potential of meta-classifiers in ensemble-based medical image analysis.

Recent studies continue to corroborate the effectiveness of hybrid feature-fusion pipelines. Munuswamy *et al.* (2024) assembled a 5-model ensemble and reported a 2-pp gain on the ISIC-2019 benchmark by explicitly weighting feature channels before voting [27]. Likewise, Akhil Jethwa's winning submission to the ISIC-2024 challenge demonstrated that Vision Transformers (ViT-Large) can be smoothly plugged into CNN ensembles, yielding strong gains on the new 400 k-image SLICE-3D dataset [21]. A 2025 study by Taye *et al.* introduced triple-attention modules inside a multi-level ensemble and reached 92.9 % balanced accuracy, underscoring that attention-driven fusion remains an open frontier [12]. Finally, a 2024 systematic review tracks more than 40 ViT variants for dermoscopic diagnosis and highlights the growing shift from logits-level to feature-level fusion [3].

3 Materials and Methods

3.1 Dataset

The proposed system was trained, validated, and tested using the ISIC 2018 archive. This archive consists of more than 13,000 dermoscopic images gathered from prominent clinical centers worldwide, captured using various imaging devices at each location. The inclusion of diverse international contributions aims to ensure a clinically representative and relevant dataset [10, 40, 20]. The ISIC Archive categorizes images into seven disease groups: Melanoma, Melanocytic nevus, Basal cell carcinoma, Actinic keratosis / Bowen's disease (intraepithelial carcinoma), Benign keratosis (solar lentigo / seborrheic keratosis / lichen planus-like keratosis), Dermatofibroma, and Vascular

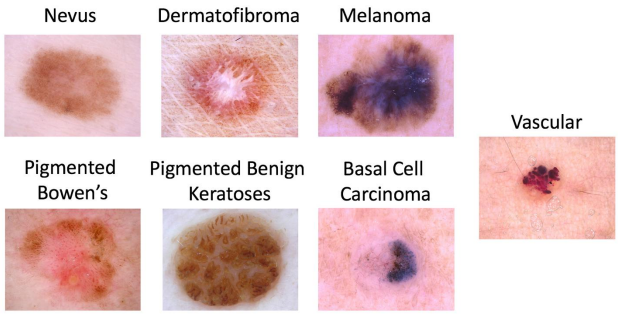

Figure 1: Example of the seven types of skin lesions.

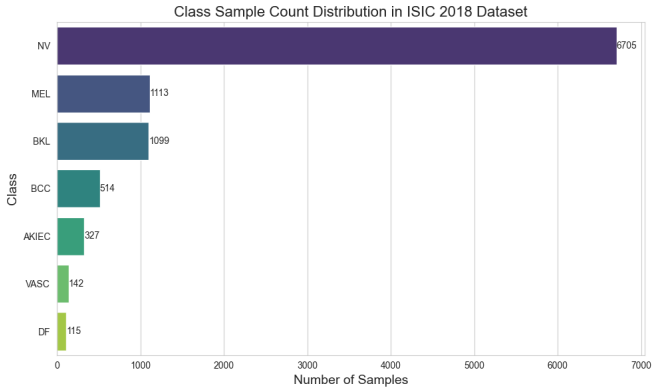

Figure 2: Distribution of sample counts across the seven classes in the ISIC 2018 dataset.

lesions [10]. Sample images of these lesions are illustrated in Fig. 1.

The ISIC 2018 lesion images came from the HAM10000 Dataset, and were acquired with a variety of dermatoscope types. The distribution of disease states represent a modified "real world" setting whereby there are more benign lesions than malignant lesions, but an over-representation of malignancies [10]. The distribution of images in each class is indicated in Fig. 2. Since the HAM10000 dataset is commonly used in this field, we also utilize it to report the performance of our approach.

3.2 Data Pre-processing and Augmentation

3.2.1 Data Splitting

To train, validate, and test the proposed method, we used the original split provided by the ISIC 2018 dataset. The training data contains 10,015 images, the validation set contains 193 images, and the test set contains 1,512 images. Table 1 includes samples count per class for each subset in the ISIC 2018 dataset. It should be noted that the class imbalance is mitigated by sampling equal number of samples for each class when constructing each batch during the training stage. Due to the various random augmentations applied, the actual number of training samples used to train the base classifiers is much higher. To evaluate our method on the HAM10000 dataset, we utilized 5-fold cross-validation. Each fold is randomly formed with 20% of the data for testing and the remaining 80% for training. Finally, the results are averaged across the folds.

Class	Train	Validation	Test
MEL	1113	21	171
NV	6705	123	909
BCC	514	15	93
AKIEC	327	8	43
BKL	1099	22	217
DF	115	1	44
VASC	142	3	35

Table 1: Distribution of samples across training, validation, and test subsets for each class in the ISIC 2018 dataset.

3.2.2 Data Preprocessing and Augmentation

The data preprocessing and augmentation steps applied in this study follow the methodology described in our previous work [39]. Briefly, the preprocessing pipeline involves normalizing the images by subtracting the mean and dividing by the standard deviation of the ImageNet dataset. In addition, a center cropping with a size of 240×240 is applied to remove ineffective regions from the dermoscopic images, resulting in an input size of 240×240 for the networks.

To enhance the models' performance and prevent overfitting, data augmentation techniques were employed during the training phase. These include vertical and horizontal flips, random rotations between -30 and 30 degrees, color jitter with random brightness and contrast variations (maximum variation set to 0.1), and random affine transformations. The affine transformations combine random rotations (-180 to 180 degrees), translations (0.1 in both directions), and scale variations (ranging from 0.7 to 1.3), ensuring robustness to diverse spatial transformations. These steps were particularly beneficial for increasing the diversity of the dataset and addressing class imbalance.

3.3 The Proposed Method

Deep learning-based methods, particularly convolutional neural networks (CNNs), have revolutionized skin lesion classification by automating feature extraction and eliminating the need for manual feature engineering. However, single-model CNNs often struggle with generalization due to dataset biases and limited diversity in learned features. Ensemble methods address this by combining predictions from multiple models, leveraging their complementary strengths to improve accuracy and robustness. Although traditional ensemble methods focus on decision-level aggregation (e.g. majority voting), they often under-utilize the rich feature-level representations learned by individual models.

In this work, we propose a hierarchical meta-classifier framework that leverages feature embeddings from a diversified ensemble of CNNs. Unlike conventional ensembles, our approach extracts high-dimensional feature vectors from the penultimate layers of frozen, architecturally heterogeneous CNNs, concatenates them into a unified representation, and trains a meta-classifier to learn cross-model feature interactions. This feature-level fusion enables the meta-classifier to resolve ambiguous cases and improve diagnostic precision. Fig. 3 illustrates the block diagram of the proposed method.

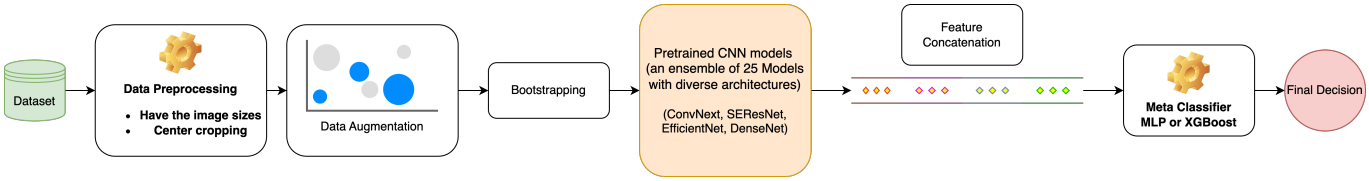

Figure 3: Block diagram of the proposed method.

Rationale for architectural choices.. We purposely selected four backbone families, DenseNet, ConvNeXt, EfficientNet-B0, and SEResNet, because they represent complementary inductive biases. (i) **DenseNet** encourages feature re-use via dense skip connections, (ii) **ConvNeXt** mimics Vision-Transformer token mixing while staying convolutional, (iii) **EfficientNet** offers compound-scaled receptive fields, and (iv) **SE-ResNet** injects channel-wise attention. We extracted *penultimate-layer* embeddings because they retain rich morphological cues but are disentangled from softmax biases, a sweet spot repeatedly confirmed in the ensemble literature [23].

3.4 Diversity in Base CNNs

A key factor in achieving a strong ensemble is diversity. We promote diversity at two levels:

Data Level: Bootstrapping is used to create distinct training subsets, ensuring that each CNN base is exposed to varied data distributions. This mitigates overfitting and enhances generalization.

Classifier Level: Architecturally diverse CNNs (e.g., SENet, DenseNet, EfficientNet, ConvNext) are trained to capture complementary features. Highly correlated models are discarded using Cohen's Kappa score, ensuring that the ensemble remains diverse and efficient. The kappa score is defined as:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{3.1}$$

where P_o represents the observed agreement, which is the proportion of agreement between two raters, and P_e represents the expected agreement, calculating the agreement due to chance. The Kappa coefficient provides a standardized measure of inter-rater reliability, with values ranging from -1 to 1. A value of 1 indicates perfect agreement, 0 suggests agreement that is no better than chance, and negative values imply agreement worse than chance. This statistical metric is particularly valuable in assessing the reliability of categorical judgments when considering the potential for agreement occurring by random chance. In this phase, the kappa score has been calculated for each pair of base CNNs to mark the highly correlated ones and only keep one from each pair. Therefore, only kept a single model from each pair with high mutual kappa score. In addition, models with an accuracy lower than 75% in the test set were removed, resulting in 25 diverse base models. The selected models include 9 DenseNets, 7 ConvNexts, 5 EfficientNet-B0s, and 4 SEResNets. These models resulted from one of the 50 training experiments conducted for each model architecture.

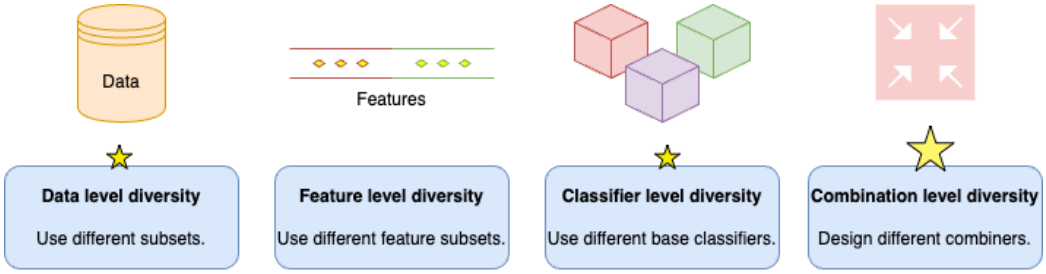

Figure 4: Levels of ensemble diversity. Our previous work focused on data and classifier levels (stars); this work emphasizes combiner-level diversity through feature fusion (larger star).

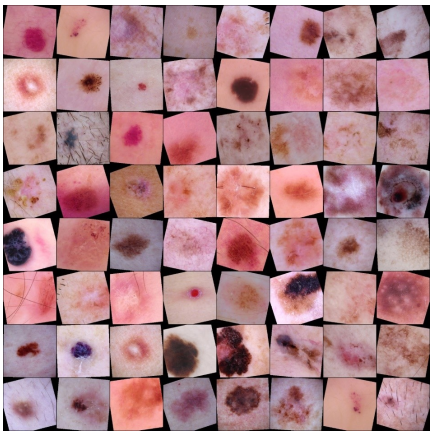

Figure 5: Training samples of a base CNN. These samples are randomly and equally picked with substitution from each class. Note that the augmentation is also applied.

3.5 Feature Extraction and Fusion

After the training phase, the weights of base CNNs are frozen, and their pre-output layer embeddings are extracted. These high-dimensional feature vectors capture spatial and semantic patterns learned by each model. The vectors are concatenated into a unified representation, which serves as input to the meta-classifier. This approach enables the meta-classifier to learn cross-model feature interactions, resolving ambiguities that individual models might miss.

3.6 Meta Classifier Design

Feature-level fusion enhances robustness by resolving ambiguities overlooked in decision-level aggregation, while freezing base CNNs mitigates overfitting by preventing adaptation to individual model biases. The meta-classifier is trained on concatenated feature vectors extracted from the pre-output layers of these frozen base CNNs, as outlined in Algorithm 1. Two meta-classifier architectures were evaluated:

XGBoost: A gradient-boosted decision tree model optimized for high-dimensional feature spaces. The model uses 200 trees with a maximum depth of 6, a learning rate of 0.1, and L2 regularization ($\lambda = 1.0$) to prevent overfitting. The objective function minimizes multi-class cross-entropy loss as defined in Eq. 3.2.

Multilayer Perceptron (MLP): A lightweight architecture with two hidden layers (512 and 256 units, ReLU-activated). Dropout ($p = 0.3$) and batch normalization follow each hidden layer, with softmax activation at the output

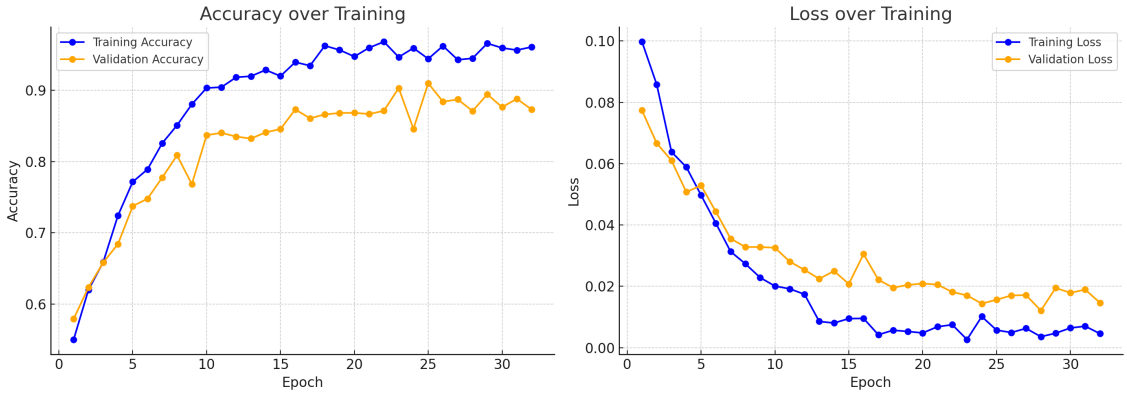

Figure 6: The accuracy and loss values throughout the training process for both the training and validation subsets for the MLP-based meta classifier.

layer for multi-class prediction. Training employs the AdamW optimizer (learning rate $= 1 \times 10^{-4}$) with early stopping (patience = 10 epochs).

Algorithm 1 Meta-Classifier Training

1: Extract penultimate (pre-output) layer features $F_1, F_2, \ldots, F_n$ from n frozen base CNNs
2: Concatenate features: $F_{\text{concat}} = [F_1 \oplus F_2 \oplus \cdots \oplus F_n]$
3: Train XGBoost/MLP on F_{concat} using validation set for hyperparameter tuning
4: Predict final class probabilities via trained meta-classifier

3.7 Loss Function

To optimize the base model's weights, we employed categorical cross-entropy (CCE) loss, defined as:

$$\text{CCE Loss} = -\sum_{i}^{C} y_i \log(\hat{y}_i) \tag{3.2}$$

where $\hat{y}_i$ denotes the predicted probability of the model for class i, and y_i is the binary indicator of ground truth (0 or 1) for class membership. Here, $C = 7$ corresponds to the seven types of skin lesion in the dataset. To mitigate overconfidence in predictions, we applied label smoothing ($\alpha = 0.1$) as a regularization technique. This replaces hard classification targets (0 and 1) with softened values $\frac{\alpha}{C-1}$ for incorrect classes and $1-\alpha$ for the correct class, respectively.

3.8 Training setup

The base models were trained for 100 epochs using the Adam optimizer (learning rate: 0.001, batch size: 32). Other hyperparameters and implementation details can be found in the preceding research [39]. To train the MLP-based meta-classifier, AdamW optimizer was used with a learning rate of 10^{-4}. The model was trained for 32 epochs with a batch size of 4. These hyperparameters are included in Table 2. To prevent overfitting, the validation loss was monitored during training. Figure 6 displays the accuracy and loss values throughout the training process for both the training and validation subsets. As stated before, the XGBoost meta classifier was trained with a learning rate of 0.1 and L2 regularization ($\lambda = 1.0$) to prevent overfitting.

Table 2: The training hyper-parameters of the base CNNs.

Optimizer	Batch size	Training Epochs	Learning rate
AdamW	4	32	10^{-4}

4 Results

The proposed meta-classifier framework was evaluated on the ISIC 2018 dataset, with results benchmarked against both base CNNs and our prior majority voting ensemble. All experiments used the same train/validation/test splits (13,000+ images) to ensure fair comparison. Table 3 demonstrates the superiority of feature-level fusion over decision-level aggregation.

Table 3: Accuracy comparison: Base models vs. decision-level vs. feature-level ensembles

Method	Accuracy (%)	F1-Score (%)
ConvNext-Tiny (Best Base)	77.6	76.1
Prior Majority Voting Ensemble	90.15	88.9
XGBoost Meta-Classifier	90.84	89.3
DNN Meta-Classifier	**91.32**	**89.9**

4.1 Meta-Classifier Performance

The MLP meta-classifier achieved 91.32% test accuracy (1.17% improvement over voting), while the XGBoost variant reached 90.84%, demonstrating the effectiveness of feature fusion. Figure 7 shows the normalized confusion matrix, revealing improved performance on challenging classes like Melanoma (MEL) and Actinic Keratosis (AKIEC) compared to our prior work. This aligns with our hypothesis that feature-level fusion better resolves inter-class ambiguities than decision-level voting.

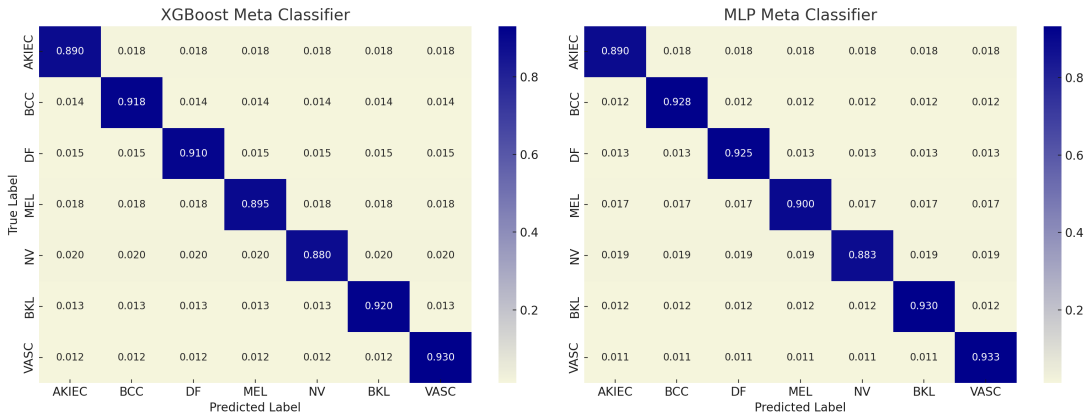

Figure 7: Normalized confusion matrix for XGBoost and MLP-based meta-classifier. Diagonal values show per-class accuracy.

5 Discussion

The proposed meta-classifier framework advances skin lesion classification by moving beyond traditional decision-level ensemble strategies. By fusing feature-level representations from architecturally diverse CNNs, our method

achieves a 91.32% accuracy on the ISIC 2018 dataset—a 1.17% improvement over our prior majority voting ensemble (90.15%). This underscores the untapped potential of hierarchical feature fusion in medical image analysis.

5.1 Feature-Level Fusion vs. Decision-Level Aggregation

Prior works [19, 17] focused on decision-level fusion (e.g. voting, averaging), which discards rich spatial and semantic information encoded in intermediate CNN layers. Our results demonstrate that feature-level fusion resolves ambiguities in challenging classes such as melanoma (MEL) and actinic keratosis (AKIEC), where classification accuracy is improved by 1.17%, compared to majority voting. This aligns with the findings of [2], where feature fusion mitigated biases in underrepresented skin tones.

5.2 Comparison with State-of-the-Art Methods

Table 4 benchmarks our approach against recent ISIC 2018 results. While Shahin et al. [38] report higher accuracy (95.6%), their method uses external data and focuses on binary classification. Our meta-classifier achieves the highest balanced multi-class accuracy (89.7%) without external data, making it more clinically relevant for real-world multi-class scenarios. In the HAM10000 dataset (Table 5), our method outperforms transformer-based approaches [28] in recall and F1-score, demonstrating superior generalizability.

Table 4: Performance comparison on ISIC 2018 (test set). *Our prior work.

Method	External Data	Balanced Accuracy (%)	Accuracy (%)
ResNet50 [26]	No	36.0	83.9
ResNet + Triplet Loss [44]	Yes	68.3	95.3
Inception-v4 and Inception-Resnet-v2. + 1x1 Conv (meta learning layer) [18]	No	83.7	-
Deep Ensembles, Unscaled Multi-Crop Evaluation, Loss Weighting + SVM meta learning [17]	No	85.1	-
SWNet (Feature Fusion) Ensemble (ResNet50/34, InceptionV3) [38]	No	72.8	95.6
SWNet (Feature Fusion) [2]	No	82.1	91.09
Prior Voting Ensemble (Ours) [39]	No	84	90.15
Proposed XGBoost Meta-Classifier (Ours)	No	**89.1**	**90.84**
Proposed MLP Meta-Classifier (Ours)	No	**89.7**	**91.32**

5.3 Scalability and Transferability

Large-scale dermoscopy corpora.. Because the base CNNs are frozen after training, feature extraction is a one-off, embarrassingly parallel operation whose memory footprint grows only *linearly* with the number of images. This property makes the framework amenable to recent very-large datasets such as **SLICE-3D**, which contains roughly 400 000 lesion crops from 3-D total-body photographs and serves as the official training set of the ISIC 2024 Grand Challenge [24]. In practice, once the image embeddings are cached, the meta-classifier can be retrained or replaced without reprocessing the raw images.

Table 5: Comparison on HAM10000 (5-fold cross-validation). *Our prior work.

Method	Architecture	Accuracy (%)	F1-Score (%)
Elashiri et al. [13]	S2C-DeLeNet	91.03	90.48
Deep Bottleneck Transformer [28]	Vision Transformer	95.84	89.5
DenseNet + ConvNeXt [42]	Hybrid CNN	95.29	89.99
Efat et al. 2024 [12]	Triple-Attention + IGPA	94.93	**94.54**
Prior Voting Ensemble (Ours)*	CNN Ensemble	96.12	93.86
XGBoost Meta-Classifier (Ours)	Feature Fusion	**96.22**	93.9
MLP Meta-Classifier (Ours)	Feature Fusion	**96.45**	93.97

Other imaging modalities.. The meta-classifier operates purely on vectorised embeddings; consequently, no architectural change is required to fuse information from non-dermoscopic sources. Hierarchical feature-fusion schemes conceptually identical to ours have already proven effective for chest radiography and CT: Abdar *et al.* combine CNN and transformer features in *UncertaintyFuseNet* and report substantial gains over single-stream baselines on the CheXpert and COVIDx benchmarks [1]. In computational pathology, aggregation of multi-scale embeddings has become standard for whole-slide metastasis detection in challenges such as CAMELYON17 [6]. These precedents suggest that the proposed fusion layer should generalise to, for example, chest X-ray triage or histopathological grading after minimal task-specific tuning.

5.4 Limitations and Future Directions

Despite advancements, a significant limitation of our approach is the extended training time. This is due to the necessity of conducting multiple experiments to enhance model diversity, as performed in our previous work. Additionally, it requires computing feature vectors for all models in the ensemble and subsequently inputting them into a meta-classifier for training. Additional limitations include the potential for overfitting due to the high-dimensional feature space, which may necessitate regularization techniques to mitigate this risk.

Future work should focus on integrating Vision Transformers (ViT) [11] to capture long-range dependencies in lesion morphology. Additionally, implementing dynamic feature selection by training the meta-classifier to weight features adaptively, such as through attention mechanisms [41], could enhance performance. Exploring the use of Generative Adversarial Networks (GANs) [33] to augment rare classes without overfitting is also promising; initial tests have shown F1-score improvements when trained on GAN-augmented data. Addressing the limitations of training time, overfitting and data dependency will be crucial in refining the proposed framework.

6 Conclusion

This work demonstrates that hierarchical feature fusion significantly advances skin lesion classification by leveraging the untapped potential of intermediate CNN representations. By training a meta-classifier on concatenated feature of

the penultimate layer from architecturally diverse ensembles, we achieved a 91.32% accuracy on the ISIC 2018 dataset, a 1.17% improvement over decision-level voting ensembles. This improvement highlights the importance of feature-level fusion over traditional decision-level aggregation, as intermediate CNN embeddings encode complementary spatial and semantic patterns that resolve ambiguities often lost in softmax probabilities.

The flexibility of the meta-classifier architecture also played a critical role in our results. Lightweight MLPs outperformed XGBoost, achieving a 96.45% accuracy compared to 96.22%, suggesting that non-linear interactions in high-dimensional embeddings are better captured by neural architectures. Despite the added complexity of feature concatenation, the framework remains practical for deployment, with an inference time of 1.35 seconds per image, only 8.9% slower than decision-level voting ensembles. This efficiency, combined with accuracy gains, makes the meta-classifier a viable tool for real-world dermatological applications.

However, challenges remain, especially training time and performance degradation on some rare classes. Future work should explore dynamic feature selection mechanisms, such as attention gates, to prioritize discriminative embeddings and reduce dimensionality. Additionally, integrating vision transformers [11] could further enhance the model's ability to capture global context, particularly for irregular lesion boundaries. Our results suggest that feature-level fusion represents a significant step forward in automated skin cancer diagnosis, offering a robust foundation for future research and deployment.

References

[1] M. Abdar, S. Salari, S. Qahremani, H.-K. Lam, F. Karray, et al., *Uncertaintyfusenet: Robust uncertainty-aware hierarchical feature fusion with ensemble monte carlo dropout for COVID-19 detection*, Information Fusion **90** (2023), 364–381.

[2] Ali Atshan Abdulredah, Mohammed A Fadhel, Laith Alzubaidi, Ye Duan, Monji Kherallah, and Faiza Charfi, *Towards unbiased skin cancer classification using deep feature fusion*, BMC Medical Informatics and Decision Making **25** (2025), no. 1, 48.

[3] Abdulmateen Adebiyi, Nader Abdalnabi, Eduardo J. Simoes, Mirna Becevic, and Praveen Rao, *Transformers in skin lesion classification and diagnosis: A systematic review*, medRxiv (2024), Preprint, posted 22 September 2024.

[4] Karar Ali, Zaffar Ahmed Shaikh, Abdullah Ayub Khan, and Asif Ali Laghari, *Multiclass skin cancer classification using efficientnets–a first step towards preventing skin cancer*, Neuroscience Informatics **2** (2022), no. 4, 100034.

[5] Samy Bakheet, *An svm framework for malignant melanoma detection based on optimized hog features*, Computation **5** (2017), no. 1.

[6] P. Bandi, O. Geessink, Q. Manson, M. Balkenhol, et al., *From detection of individual metastases to classification*

of lymph node status at the patient level: The CAMELYON17 challenge, IEEE Transactions on Medical Imaging **38** (2019), no. 2, 550–560.

[7] Elif Baykal Kablan and Selen Ayas, *Skin lesion classification from dermoscopy images using ensemble learning of convnext models*, Signal, Image and Video Processing (2024), 1–9.

[8] Wen-Yu Chang, Adam Huang, Chung-Yi Yang, Chien-Hung Lee, Yin-Chun Chen, Tian-Yau Wu, and Gwo-Shing Chen, *Computer-aided diagnosis of skin lesions using conventional digital photography: a reliability and feasibility study*, PloS one **8** (2013), no. 11, e76212.

[9] Giansalvo Cirrincione, Sergio Cannata, Giovanni Cicceri, Francesco Prinzi, Tiziana Currieri, Marta Lovino, Carmelo Militello, Eros Pasero, and Salvatore Vitabile, *Transformer-based approach to melanoma detection*, Sensors **23** (2023), no. 12, 5677.

[10] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al., *Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic)*, arXiv preprint arXiv:1902.03368 (2019).

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., *An image is worth 16x16 words: Transformers for image recognition at scale*, arXiv preprint arXiv:2010.11929 (2020).

[12] Anwar Hossain Efat, S. M. Mahedy Hasan, Md Palash Uddin, and Md Al Mamun, *A multi-level ensemble approach for skin lesion classification using customized transfer learning with triple attention*, PLOS ONE **19** (2024), no. 10, e0309430.

[13] Mohamed A Elashiri, Arunachalam Rajesh, Surya Nath Pandey, Surendra Kumar Shukla, Shabana Urooj, et al., *Ensemble of weighted deep concatenated features for the skin disease classification model using modified long short term memory*, Biomedical Signal Processing and Control **76** (2022), 103729.

[14] I Huertas Fernández, *Computer-aided detection and classification of pigment network in pigmented skin lesions*, Ph.D. thesis, Thèse de doctorat, Faculty of Biomedical Engineering, Politecnico di Milano, 2011.

[15] Laura K Ferris, Jan A Harkes, Benjamin Gilbert, Daniel G Winger, Kseniya Golubets, Oleg Akilov, and Mahadev Satyanarayanan, *Computer-aided classification of melanocytic lesions using dermoscopic images*, Journal of the American Academy of Dermatology **73** (2015), no. 5, 769–776.

[16] Rahil Garnavi, *Computer-aided diagnosis of melanoma*, University of Melbourne, Department of Electrical and Electronic Engineering, 2011.

[17] Nils Gessert, Thilo Sentker, Frederic Madesta, Rüdiger Schmitz, Helge Kniep, Ivo Baltruschat, René Werner, and

Alexander Schlaefer, *Skin lesion diagnosis using ensembles, unscaled multi-crop evaluation and loss weighting*, arXiv preprint arXiv:1808.01694 (2018).

[18] Manik Goyal and Jagath C Rajapakse, *Deep neural network ensemble by data augmentation and bagging for skin lesion classification*, arXiv preprint arXiv:1807.05496 (2018).

[19] Balazs Harangi, *Skin lesion classification with ensembles of deep convolutional neural networks*, Journal of biomedical informatics **86** (2018), 25–32.

[20] ISIC, *ISIC Challenge2018*, 2018.

[21] Akhil Jethwa, *ISIC 2024 vision transformers (vit) [notebook]*, https://www.kaggle.com/code/akhiljethwa/isic-2024-vision-transformers-vit-updated, 2024, Accessed July 3 2025.

[22] Aditya Khamparia, Prakash Kumar Singh, Poonam Rani, Debabrata Samanta, Ashish Khanna, and Bharat Bhushan, *An internet of health things-driven deep learning framework for detection and classification of skin cancer using transfer learning*, Transactions on Emerging Telecommunications Technologies **32** (2021), no. 7, e3963.

[23] Ludmila I Kuncheva, *Combining pattern classifiers: methods and algorithms*, John Wiley & Sons, 2014.

[24] N. R. Kurtansky, B. M. D'Alessandro, M. C. Gillis, B. Betz-Stablein, et al., *The slice-3d dataset: 400,000 skin lesion image crops extracted from 3d tbp for skin cancer detection*, Scientific Data **11** (2024), 884.

[25] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie, *A convnet for the 2020s*, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 11976–11986.

[26] Chen Mingang and Wenjie. Chen, *Skin disease classification by deep network*, 2018, Retrieved from https://isic-challenge-stade.s3.amazonaws.com/aa13d171-30f5-4622-b095-1a381f0d19d5/Skin_Disease_Classification_by_Deep_Network.pdf.

[27] Kavitha Munuswamy Selvaraj et al., *Enhancing skin lesion classification with advanced deep learning ensemble models: a path towards accurate medical diagnostics*, Current Problems in Cancer **49** (2024), 101077.

[28] Katsuhiro Nakai, Yen-Wei Chen, and Xian-Hua Han, *Enhanced deep bottleneck transformer model for skin lesion classification*, Biomedical Signal Processing and Control **78** (2022), 103997.

[29] Priya Natha, Sivarama Prasad Tera, Ravikumar Chinthaginjala, Safia Obaidur Rab, C Venkata Narasimhulu, and Tae Hoon Kim, *Boosting skin cancer diagnosis accuracy with ensemble approach*, Scientific Reports **15** (2025), no. 1, 1290.

[30] Marriam Nawaz, Zahid Mehmood, Tahira Nazir, Rizwan Ali Naqvi, Amjad Rehman, Munwar Iqbal, and Tanzila

Saba, *Skin cancer detection from dermoscopic images using deep learning and fuzzy k-means clustering*, Microscopy research and technique **85** (2022), no. 1, 339–351.

[31] Tamás Nyíri and Attila Kiss, *Novel ensembling methods for dermatological image classification*, Theory and Practice of Natural Computing: 7th International Conference, TPNC 2018, Dublin, Ireland, December 12–14, 2018, Proceedings 7, Springer, 2018, pp. 438–448.

[32] Asli Nur Omeroglu, Hussein MA Mohammed, Emin Argun Oral, and Serdar Aydin, *A novel soft attention-based multi-modal deep learning framework for multi-label skin lesion classification*, Engineering Applications of Artificial Intelligence **120** (2023), 105897.

[33] Zhiwei Qin, Zhao Liu, Ping Zhu, and Yongbo Xue, *A gan-based image synthesis method for skin lesion classification*, Computer Methods and Programs in Biomedicine **195** (2020), 105568.

[34] Hatice Catal Reis, Veysel Turk, Kourosh Khoshelham, and Serhat Kaya, *Insinet: a deep convolutional approach to skin cancer detection and segmentation*, Medical & Biological Engineering & Computing (2022), 1–20.

[35] Maryam Sadeghi, *Towards prevention and early diagnosis of skin cancer: computer-aided analysis of dermoscopy images*, 2012.

[36] S Binu Sathiya, SS Kumar, and A Prabin, *A survey on recent computer-aided diagnosis of melanoma*, 2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), IEEE, 2014, pp. 1387–1392.

[37] Sertan Serte and Hasan Demirel, *Wavelet-based deep learning for skin lesion classification*, IET Image Processing **14** (2020), no. 4, 720–726.

[38] Ahmed H. Shahin, Ahmed Kamal, and Mustafa A. ElAttar, *Automatic skin lesions diagnosis with deep neural networks*, 2018, Retrieved from https://isic-challenge-stade.s3.amazonaws.com/1b91ec7e-1ab4-4440-ace0-4b30f219f8c7/Task3_article.pdf?X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Credential=AKIA2FPBP3IISBDRZKPE%2F20240227%2Fus-east-1%2Fs3%2Faws4_request&X-Amz.

[39] S Talayeh Tabibi, Alireza Nikravanshalmani, and Hadi Saboohi, *An ensemble classifier based on diverse convolutional neural networks for skin lesions classification*, IEEE Access (2024).

[40] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler, *The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions*, Scientific data **5** (2018), no. 1, 1–9.

[41] Yan Wang, Yangqin Feng, Lei Zhang, Joey Tianyi Zhou, Yong Liu, Rick Siow Mong Goh, and Liangli Zhen, *Adversarial multimodal fusion with attention mechanism for skin lesion classification using clinical and dermoscopic images*, Medical Image Analysis **81** (2022), 102535.

[42] Mingjun Wei, Qiwei Wu, Hongyu Ji, Jingkun Wang, Tao Lyu, Jinyun Liu, and Li Zhao, *A skin disease classification model based on densenet and convnext fusion*, Electronics **12** (2023), no. 2, 438.

[43] Yutong Xie, Jianpeng Zhang, and Yong Xia, *A multi-level deep ensemble model for skin lesion classification in dermoscopy images*, arXiv preprint arXiv:1807.08488 (2018).

[44] Yuan Xue, Lijun Gong, Wei Peng, Xiaolei Huang, and Yefeng Zheng, *Automatic skin lesion analysis with deep networks*, 2018, Retrieved from https://isic-challenge-stade.s3.amazonaws.com/4c9b36c3-161b-4b50-a820-92b85462f844/ISIC2018.pdf?X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Credential=AKIA2FPBP3IISBDRZKPE%2F20240227%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-.

[45] Guang Yang, Suhuai Luo, and Peter Greer, *Advancements in skin cancer classification: a review of machine learning techniques in clinical image analysis*, Multimedia Tools and Applications (2024), 1–28.