Entropy-Aware Skin Lesion Classification via Conformal Ensemble of Vision Transformers

by

Mehran Zoravar
B.Eng., Amirkabir University of Technology, 2022

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF APPLIED SCIENCE

in the Department of Mechanical Engineering

We acknowledge and respect the Lək̓ʷəŋen (Songhees and Xʷsepsəm/Esquimalt) Peoples on whose territory the university stands, and the Lək̓ʷəŋen and W̱SÁNEĆ Peoples whose historical relationships with the land continue to this day.

Entropy-Aware Skin Lesion Classification via Conformal Ensemble of Vision Transformers

by

Mehran Zoravar
B.Eng., Amirkabir University of Technology, 2022

Supervisory Committee

Dr. Homayoun Najjaran, Supervisor
(Department of Mechanical Engineering, University of Victoria)

Dr. Flavio Firmani, Departmental Member
(Department of Mechanical Engineering, University of Victoria)

ABSTRACT

Uncertainty quantification is an inherent part of decision-making in various domains, lending credibility and interpretability to predictive models. In medical image analysis, where decisions involve high stakes, the precise prediction becomes an imperative task. Skin lesion classification is a significant application in this domain, necessitating robust uncertainty handling to ensure diagnostic accuracy. This paper presents the Entropy-Aware Conformal Ensemble of Vision Transformers (EACE-ViTs), which applied Vision Transformer (ViT) models, Generative Adversarial Networks (GANs) for efficient data augmentation, and entropy-based ensemble weightings in a synergistic manner to meet these requirements. The proposed framework addresses key challenges in the domain, including class imbalance and the prevalent problem of low-confidence predictions due to the intricate nature of skin lesions. A distinctive feature of this work is the adaptive entropy adjustment applied to the thresholds in Regularized Adaptive Prediction Sets (RAPS) and Adaptive Prediction Sets (APS) methods, which calibrates uncertainty based on model confidence. This kind of adjustment provides more adaptive and flexible prediction sets. When tested with the HAM10000 dataset, EACE-ViTs not only exhibits better performance in accuracy, coverage, and uncertainty metrics but also outperforms the baseline CNN-based models as well as individual ViT models by a considerable margin. The results validate the potential of EACE-ViTs as a promising tool for accurate and reliable classification of skin lesions in important medical applications..

Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

I would like to thank:

Dr. Homayoun Najjaran, for his invaluable guidance, mentorship, encouragement, and unwavering support throughout my research journey.

My colleagues at the Advanced Control and Intelligence System (ACIS) lab, for their insightful discussions, collaboration, and support that enriched this research.

The University of Victoria, for providing the resources and an excellent academic environment that made this work possible.

My family and friends, for their unconditional love, encouragement, and patience, which have been a constant source of strength.

DEDICATION

Just hoping this is useful!

Chapter 1

Introduction

1.1 Introduction

The objective of this thesis is to introduce a novel deep learning framework, named **Entropy-Aware Conformal Ensemble of Vision Transformers (EACE-ViTs)**, for the classification of skin lesions in dermoscopic images. The proposed framework is designed to tackle multiple persistent challenges in medical image analysis, with a focus on improving predictive accuracy, handling class imbalance, and incorporating reliable uncertainty quantification. In recent years, the application of artificial intelligence in the medical domain has expanded significantly, enabling more accurate, efficient, and scalable diagnostic solutions. However, as these models begin to transition from research environments into real-world clinical workflows, their reliability, interpretability, and ability to manage uncertainty become paramount. In high-stakes scenarios such as skin cancer detection, where the consequences of misclassification may include delayed diagnosis or over treatment, it is not sufficient for a model to simply be accurate—it must also be transparent, robust, and able to communicate the confidence of its predictions.

Skin lesion classification represents a particularly demanding task within the broader field of computer-aided diagnosis. Skin cancer, especially malignant melanoma, is one of the most common forms of cancer worldwide, and its incidence continues

to rise. Dermoscopy, a non-invasive imaging technique that enhances visualization of skin surface structures, has become the clinical standard for evaluating suspicious pigmented lesions. Yet even with dermoscopic assistance, diagnostic accuracy can vary substantially between experts. This inconsistency has motivated the development of automated image analysis systems, which aim to support dermatologists by offering consistent and objective second opinions. Deep learning approaches, particularly those based on convolutional neural networks (CNNs), have achieved impressive performance in this domain. More recently, Vision Transformers (ViTs) have emerged as a powerful alternative, demonstrating state-of-the-art results across various image recognition benchmarks, including medical imaging tasks.

ViTs offer several advantages over traditional CNNs, especially in their ability to capture long-range dependencies and global contextual features within an image. Unlike CNNs, which rely on localized convolutional filters and pooling operations, Vision Transformers employ self-attention mechanisms to model the relationships between all parts of an image. This property is especially useful in dermatology, where lesion features such as symmetry, color distribution, and border irregularities are dispersed across the entire image and must be interpreted holistically. Despite their promise, however, ViT-based models face several critical limitations that hinder their deployment in clinical practice. Notably, most models provide deterministic outputs without any quantification of uncertainty, making it difficult for healthcare providers to gauge the reliability of each prediction. Moreover, class imbalance in dermatology datasets such as HAM10000 further complicates model performance, particularly for underrepresented but clinically significant lesion types.

To address these challenges, this thesis proposes a hybrid framework that combines the strengths of Vision Transformers with ensemble learning, conformal prediction, and entropy-based uncertainty modeling. The resulting system, EACE-ViTs, is designed not only to enhance classification performance but also to provide interpretable and statistically valid prediction sets that quantify uncertainty. Furthermore, the framework integrates a Generative Adversarial Network (GAN) for data augmentation, specifically targeting the imbalance between frequent and rare lesion classes. By generating high-quality synthetic samples, the GAN component enriches

the training set and improves the generalization ability of the model across diverse lesion types. Through this integrated approach, EACE-ViTs advances the state of the art in both predictive accuracy and clinical reliability.

This introductory chapter serves to frame the goals and scope of the thesis. In the sections that follow, we articulate the core motivations behind the development of EACE-ViTs and highlight the methodological innovations that distinguish it from prior work. We also present a concise summary of the key contributions of this research and describe the structure of the thesis. The subsequent chapters delve deeper into the related literature, technical implementation, experimental validation, and broader implications of the proposed framework. Taken together, this work aspires to contribute meaningfully to the development of intelligent diagnostic tools that are not only accurate but also safe, interpretable, and aligned with the needs of modern healthcare systems.

1.2 Motivation

Skin cancer is among the most prevalent and deadly forms of cancer worldwide, and early detection remains the most effective method for reducing morbidity and mortality. Malignant melanoma, in particular, progresses rapidly and is often curable when identified at an early stage. The advent of dermoscopy has improved the ability of clinicians to examine skin lesions in detail, but manual inspection remains subjective and prone to diagnostic variability. In this context, deep learning—especially vision-based models—has emerged as a powerful tool capable of supporting dermatologists with consistent and objective assessments.

Among the recent developments in deep learning, Vision Transformers (ViTs) have gained considerable attention for their ability to model global and long-range dependencies in images. Unlike convolutional neural networks (CNNs), which rely on fixed receptive fields, ViTs utilize self-attention mechanisms to dynamically capture both fine-grained and contextual information across the entire image. This capability is particularly advantageous in medical imaging tasks where lesion boundaries may be ambiguous and global context often plays a critical role in diagnosis. Despite their

promise, however, ViTs and other deep models face several fundamental challenges when applied to clinical domains.

First and foremost, one of the most pressing challenges in medical AI is the lack of reliable uncertainty estimation. Most conventional models provide point predictions without expressing any confidence level, making it difficult for clinicians to interpret and trust their outputs. In clinical settings, overconfident incorrect predictions can be dangerous, leading to either false reassurance or unwarranted concern. Therefore, incorporating uncertainty-aware mechanisms is crucial to align model behavior with the caution exercised by healthcare professionals.

Secondly, medical image datasets, including the widely used HAM10000 dataset, suffer from significant class imbalance. While common lesion types such as benign nevi are well represented, rarer but clinically critical classes like dermatofibromas, vascular lesions, or actinic keratoses often contain a limited number of samples. This imbalance skews the learning process, resulting in models that perform well on dominant classes but poorly on minority classes—precisely where accurate classification is most essential.

Finally, the issue of model trustworthiness extends beyond performance metrics. For medical AI to be integrated into diagnostic workflows, models must exhibit not only high accuracy but also robustness to variations in input data, interpretability in their decisions, and adaptability to novel or uncertain cases. Standard training paradigms often neglect these aspects, focusing narrowly on optimizing for accuracy while overlooking the broader requirements of safe and equitable decision-making.

Addressing these challenges requires a multifaceted approach that integrates state-of-the-art advances in deep learning, uncertainty quantification, and data augmentation. The framework proposed in this thesis seeks to meet this need by combining entropy-based filtering, conformal prediction methods, and ensemble learning to deliver high-performance classification with well-calibrated confidence intervals. Through the use of Generative Adversarial Networks (GANs), we also generate synthetic examples for underrepresented lesion classes, enhancing data diversity and model generalizability. The result is a robust diagnostic tool capable of making nuanced predictions with measurable and interpretable uncertainty—an essential prop-

erty in modern healthcare applications.

This thesis evaluates the proposed framework using the HAM10000 dataset, a widely recognized dermatoscopic image collection of 10,015 images across seven diagnostic categories. While this dataset is valuable for benchmarking automated methods, it has limitations: the distribution of classes is highly imbalanced, and the population lacks diversity in terms of skin tones, environments, and geographic representation. These factors restrict the generalizability of models trained solely on HAM10000 and motivate the need for augmentation strategies and future validation on larger, more diverse datasets.

1.3 Contributions

The work presented in this thesis makes several novel and practical contributions to the fields of medical imaging, machine learning, and uncertainty-aware prediction systems:

1. **Novel Framework:** We introduce the **EACE-ViTs** framework, which combines Vision Transformers, entropy-aware ensemble strategies, conformal prediction, and GAN-based data augmentation to improve diagnostic performance in skin lesion classification. This integrated approach represents a substantial advancement over existing methods that typically focus on a single aspect of the problem.

2. **Entropy-Aware Calibration:** We propose the use of entropy-based thresholds to dynamically adjust the conformal prediction process, allowing the model to expand or contract prediction sets based on input uncertainty. This strategy enhances the flexibility and reliability of the predictive intervals produced by APS and RAPS methods.

3. **Synthetic Data Generation:** We employ a conditional GAN model to generate high-quality synthetic images for rare lesion categories in the HAM10000 dataset. This augmentation process addresses the dataset's class imbalance

and improves the model's ability to generalize across all classes, including those with few training examples.

4. **Comprehensive Experimental Validation:** We conduct an extensive evaluation of the proposed method using the HAM10000 dataset, benchmarking it against several baselines, including CNNs, single ViTs, and ensemble models without uncertainty quantification. Our results demonstrate consistent improvements in classification accuracy, macro-level metrics, and uncertainty calibration.

5. **Clinical Relevance and Interpretability:** The proposed framework is designed with clinical applicability in mind. By offering interpretable prediction sets and dynamic confidence adjustment, EACE-ViTs aligns more closely with real-world diagnostic needs, paving the way for safer integration of AI into healthcare workflows.

1.4 Thesis Outline

This thesis is structured to guide the reader through the research process, from problem formulation to methodology, evaluation, and future implications.

Chapter 1: Introduces the motivation, research challenges, contributions, and structural organization of the thesis. It frames the research problem in the context of clinical AI deployment and highlights the importance of uncertainty-aware predictions in medical imaging.

Chapter 2: Reviews the existing literature related to skin lesion classification, Vision Transformers, ensemble learning, uncertainty quantification, and conformal prediction. It identifies gaps in current approaches and situates the proposed framework within the broader research landscape.

Chapter 3: Describes the design and implementation of the **EACE-ViTs** framework, including data preprocessing, synthetic image generation, entropy-based

ensemble filtering, and conformal prediction integration. It details each module's architecture and the rationale behind its inclusion.

Chapter 4: Presents the experimental setup, evaluation metrics, and results obtained from testing the proposed framework. This chapter includes both quantitative and qualitative analyses, along with ablation studies to understand the impact of each component.

Chapter 5: Summarizes the findings, reiterates the contributions, and outlines directions for future work. This chapter reflects on the broader implications of this research and its potential extension to other medical domains.

Through this structure, the thesis offers a comprehensive and coherent exploration of entropy-aware conformal prediction in skin lesion classification. Each chapter builds upon the previous one to tell a complete story—one that begins with a clinical need and ends with a practical, validated, and deployable machine learning solution.

Chapter 2

Skin Lesion Classification and Literature Review

In this chapter, the challenges and motivations behind accurate and reliable skin lesion classification are presented. This chapter begins with an overview of the context and challenges in medical image analysis, particularly focusing on skin lesion classification. It then reviews existing approaches and related works in the field, highlighting their limitations and areas for improvement. Finally, it positions the research contributions of this thesis within the broader landscape of medical imaging research.

2.1 Introduction to the Problem

Recent advancements in deep learning have revolutionized medical image analysis, especially in tasks like skin lesion classification. Skin lesions, particularly those indicative of melanoma, pose a unique challenge due to their high intra-class variability and inter-class similarity, making robust classification essential for accurate diagnosis. Traditionally, Convolutional Neural Networks (CNNs) have dominated the field of medical imaging, leveraging their localized receptive fields to capture visual patterns effectively [21]. However, CNNs struggle to capture long-range dependencies

within images, limiting their utility in tasks where global context is critical, such as distinguishing subtle variations in skin lesion textures and shapes.

The introduction of the Vision Transformer (ViT) by Dosovitskiy et al. marked a transformative step in deep learning for vision tasks, with the model's self-attention mechanism enabling it to capture both local and global features effectively [14]. This characteristic has shown promise in medical imaging, where a nuanced understanding of image-wide context is vital [3]. For instance, recent studies like DeepSkin by Gururaj et al. demonstrate the effectiveness of deep learning in skin lesion classification, leveraging large annotated datasets like HAM10000 to improve model accuracy [20]. As ViTs gain traction in medical imaging, variants such as SkinDistilViT have shown that integrating ViTs with ensemble learning can enhance classification performance while maintaining model robustness [29].

While achieving high accuracy is critical, recent research emphasizes the importance of model trustworthiness in clinical applications, where incorrect predictions can have severe consequences. Traditional metrics like accuracy fall short of assessing the reliability of predictions in real-world scenarios. In response, uncertainty quantification (UQ) has emerged as a crucial component, providing additional insights into the model's confidence in its predictions. Conformal prediction, a statistical framework for quantifying uncertainty, has demonstrated particular effectiveness in high-stakes applications, offering prediction intervals that maintain a specified level of coverage probability [6]. Integrating conformal prediction with Vision Transformers can thus enable entropy-aware models that dynamically adjust their predictions based on uncertainty, an essential attribute in medical diagnostics.

This paper proposes a novel approach that uses Vision Transformers, ensemble learning, and conformal prediction to achieve entropy-aware skin lesion classification. By utilizing conformal ensembles of ViTs, our model not only enhances predictive accuracy but also provides calibrated prediction intervals that quantify uncertainty. This framework allows for a more robust and reliable model suitable for clinical applications, where both high accuracy and dependable uncertainty estimates are paramount. This work contributes to the ongoing research on robust AI systems in healthcare, aiming to provide a framework that can support clinicians in making

informed and reliable diagnostic decisions.

2.2 Related Work

2.2.1 Skin Lesion Classification

Skin lesion classification is a critical area of research in dermatology, where early detection of malignant lesions like melanoma can significantly improve patient outcomes. With the advent of deep learning, various approaches have been developed to enhance the accuracy and reliability of skin lesion classification systems. Convolutional Neural Networks (CNNs) have been widely used for this task, but more recent advancements, such as Vision Transformers (ViTs), have shown promising results by capturing more global context and fine-grained details compared to traditional CNNs [14].

A comprehensive review by [38] presents a review of deep learning-based models specifically applied to skin lesion diagnosis, emphasizing the unique challenges posed by skin lesion datasets, such as inter-class similarities and intra-class variability. To address these challenges, domain adaptation techniques have also been explored to improve model generalizability across diverse datasets. For instance, [19] discusses the effectiveness of domain adaptation in enhancing skin lesion classification, particularly when models are trained on one dataset and tested on another, thus mitigating data distribution shifts.

In addition to model architecture improvements, there has been an increased focus on understanding and quantifying the uncertainty in skin lesion predictions. Uncertainty quantification techniques, as introduced by [6], provide additional safety margins by indicating the confidence of predictions, which is particularly valuable in clinical settings. Furthermore, [2] has examined Vision Transformers specifically within the context of skin lesion classification, highlighting their superior performance on datasets like ISIC, a widely recognized dataset for skin lesion analysis [23].

Lastly, empirical studies such as [16] have evaluated the performance of various

deep learning models for skin cancer detection, underscoring the importance of robust and reliable models in clinical practice. By leveraging advanced architectures and incorporating uncertainty measures, these models hold potential for assisting dermatologists in more accurately diagnosing and classifying skin lesions.

2.2.2 Ensemble Learning in Medical Imaging

Ensemble learning has gained significant attention in machine learning due to its ability to improve prediction accuracy by combining the strengths of multiple models. Early work by [9] introduced "bagging," which leverages the diversity of models generated from bootstrap samples to reduce variance and improve robustness in predictions. [13] further explored the theoretical benefits of ensemble methods, noting their effectiveness in overcoming the limitations of individual classifiers through diverse combinations.

In medical imaging, ensemble methods have shown substantial promise due to their ability to handle complex and noisy data. For instance, recent advances highlight that ensemble deep learning can enhance diagnostic accuracy by combining different neural network architectures and training techniques, as emphasized by [38]. Ensemble approaches, including stacking, boosting, and bagging, are famous for integrating predictions from multiple deep learning models, thereby capturing a broader range of features that might be missed by a single model. [42] introduced stacked generalization, which combines base learners with a meta-learner and remains a foundational approach for improving predictive accuracy in such complex tasks.

Boosting algorithms like [17]'s Adaboost and [18]'s gradient boosting have demonstrated notable success in incrementally focusing on misclassified instances, making them particularly effective for enhancing sensitivity in anomaly detection and rare class identification in medical data. The application of ensemble learning in medical imaging is further enriched by ensemble deep learning techniques that integrate various neural network models, as reviewed by [33] and [43], showcasing the potential for optimized decision-making in clinical settings. These advances emphasize the

adaptability and scalability of ensemble methods, making them well-suited to the demanding requirements of medical imaging. While the preceding overview highlights the importance and overall impact of ensemble methods in medical imaging, a deeper understanding of the specific techniques is essential for comprehending their unique strengths and applications. In this section, we delve into the detailed mechanisms, advantages, and challenges of prominent ensemble learning techniques, starting with Bagging.

2.2.2.1 Bagging

Bagging, short for **Bootstrap Aggregating**, introduced by Breiman [9], is a method that enhances prediction accuracy by combining multiple versions of a model trained on different subsets of data. The subsets are generated through **bootstrap sampling**, where random samples are drawn with replacement from the original dataset.

The core idea of bagging is to reduce the variance of a model by aggregating predictions from multiple base models. Given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ with N samples, bagging creates an ensemble of models $\{M_k\}_{k=1}^{K}$, each trained on a different bootstrap sample $\mathcal{D}_k$. A bootstrap sample is generated by randomly selecting N instances from $\mathcal{D}$ with replacement, meaning that some instances may appear multiple times, while others may be omitted.

Once the base models are trained, their predictions are aggregated during the prediction phase. For regression tasks, the final prediction for an input x is obtained by averaging the predictions from all models:

$$\hat{y} = \frac{1}{K} \sum_{k=1}^{K} M_k(x)$$

For classification tasks, a majority voting scheme is used, where each model votes for a class, and the class with the highest number of votes is selected as the final

prediction:

$$\hat{y} = \arg\max_j \sum_{k=1}^{K} I(M_k(x) = j)$$

where $I(\cdot)$ is an indicator function that equals 1 if the condition is true and 0 otherwise.

Bagging works particularly well for models that are prone to high variance, such as decision trees and neural networks. By aggregating multiple predictions, bagging reduces the impact of any single model's error, leading to more stable and accurate predictions. The method is less effective for stable models, such as k-nearest neighbors, where variance is already low.

In practice, bagging can be implemented efficiently and is inherently parallelizable since each model can be trained independently. One important consideration is the number of bootstrap samples K, which controls the size of the ensemble. Although higher values of K typically improve performance, they also increase computational cost. In most applications, a value of K between 25 and 50 is sufficient to achieve significant variance reduction.

Despite its advantages, bagging has limitations. It increases computational cost because multiple models must be trained and aggregated. Moreover, interpretability is reduced, particularly when complex models are used as base learners.

In summary, bagging is a robust ensemble method that enhances prediction accuracy by reducing variance. Its simplicity and effectiveness make it a fundamental technique in machine learning, especially for high-variance models. Breiman's original work laid the foundation for numerous subsequent developments in ensemble learning, demonstrating that bagging can significantly improve both regression and classification tasks in diverse real-world applications.

2.2.2.2 Boosting

Boosting is an ensemble learning method designed to improve the performance of weak learners by combining them into a single strong model. A weak learner is a model that performs slightly better than random guessing. The foundational frame-

work for boosting was introduced by Freund and Schapire in their decision-theoretic generalization of online learning, leading to the development of the well-known AdaBoost algorithm.

The core idea of boosting is to train multiple weak learners sequentially, where each learner focuses on the mistakes made by its predecessors. This process ensures that subsequent models improve on the areas where earlier models performed poorly, thereby incrementally enhancing the overall accuracy of the ensemble.

Boosting begins by assigning equal weights to all training instances. In the first iteration, a weak learner is trained on the dataset, and its error rate is calculated. The error rate is the proportion of misclassified instances. The algorithm then updates the weights of the training instances: instances that were misclassified by the weak learner receive higher weights, meaning that they will have more influence in the next iteration. Conversely, correctly classified instances receive lower weights.

In each subsequent iteration, a new weak learner is trained on the weighted dataset. The process of updating the weights ensures that the new learner focuses more on the difficult cases that previous learners struggled to classify. The weight assigned to each learner's prediction is determined based on its accuracy; learners with lower error rates are given higher weights in the final prediction.

Mathematically, the weight update for each training instance w_i after iteration t is performed using the following rule:

$$w_i \leftarrow w_i \cdot \exp(\alpha_t \cdot I(h_t(x_i) \neq y_i))$$

where α_t is the weight of the weak learner h_t, computed as:

$$\alpha_t = \frac{1}{2} \ln \left(\frac{1 - \epsilon_t}{\epsilon_t} \right)$$

and ϵ_t is the error rate of the weak learner h_t. The indicator function $I(\cdot)$ returns 1 if the prediction $h_t(x_i)$ is incorrect, and 0 otherwise.

After each iteration, the weights are normalized to ensure they sum to 1:

$$w_i \leftarrow \frac{w_i}{\sum_{j=1}^{N} w_j}$$

This iterative process continues for a predetermined number of iterations T or until the ensemble reaches a desired level of accuracy. At the end of the training process, the final strong classifier $H(x)$ is obtained by taking a weighted majority vote of all the weak learners:

$$H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t \cdot h_t(x)\right)$$

In boosting, each weak learner contributes to the final prediction according to its accuracy, ensuring that more reliable learners have a greater influence on the ensemble's output. This approach reduces both bias and variance, making boosting a powerful method for improving model performance.

Boosting is particularly effective in cases where weak learners can be incrementally improved, such as decision trees with a small depth (commonly referred to as decision stumps). However, one limitation of boosting is its sensitivity to noisy data and outliers, as the algorithm may focus excessively on difficult instances that could be anomalies.

In summary, boosting is a sophisticated ensemble method that converts weak learners into a strong predictor by iteratively focusing on challenging instances and aggregating the results. Its success in a variety of machine learning tasks has made it a fundamental tool in both theoretical and practical applications.

2.2.2.3 Stacking

Stacking is a powerful ensemble learning technique that differs from methods like bagging and boosting in its approach to combining multiple models. Unlike bagging, which reduces variance by averaging predictions from base models, and boosting, which iteratively improves weak models by focusing on difficult examples, stacking

aims to leverage the strengths of various base models by training a meta-learner on their outputs. The goal is to create a more robust and accurate final model by allowing the meta-learner to identify patterns in the predictions made by the base models.

The general process of stacking involves two key stages: training base models and training a meta-learner. Initially, several different base models, often referred to as level-0 models, are trained on the same dataset. These base models can be of different types, such as decision trees, neural networks, or support vector machines, ensuring a diverse set of predictive capabilities. Once the base models are trained, their predictions are collected to form a new dataset, where each feature corresponds to the prediction of a base model.

Next, a meta-learner, also known as a level-1 model, is trained on this new dataset. The meta-learner's role is to learn how to best combine the outputs of the base models to produce a final prediction. The choice of the meta-learner can vary depending on the problem and the nature of the base models; common choices include linear regression for regression tasks and logistic regression for classification tasks. The meta-learner typically uses k-fold cross-validation on the training set to generate unbiased predictions from the base models.

One of the main advantages of stacking is its ability to combine models with different inductive biases, potentially capturing a broader range of patterns in the data. This makes stacking particularly effective in complex tasks where no single model performs optimally across all regions of the input space. However, the method also introduces additional computational complexity due to the need for training multiple models and performing cross-validation.

The concept of a meta-learner in stacking is closely related to meta-learning, which refers to the process of learning how to learn. In the context of stacking, the meta-learner operates at a higher level of abstraction, learning from the predictions of base models rather than directly from the raw data. This hierarchical approach allows stacking to exploit relationships between the outputs of different models, potentially improving generalization performance.

The effectiveness of stacking has been demonstrated in numerous studies and

practical applications. For example, research by Wolpert (1992) [42]introduced the original concept of stacked generalization, showing its potential to improve predictive accuracy. Subsequent work has explored various extensions and modifications, such as using different types of meta-learners and incorporating additional features alongside the base models' predictions.

Despite its benefits, stacking has certain limitations. It requires careful selection of base models and the meta-learner to avoid overfitting, especially when the training data is limited. Moreover, interpretability can be challenging, as the final prediction depends on the combined outputs of multiple models. Nevertheless, when appropriately configured, stacking remains a highly effective ensemble technique, particularly in scenarios where diverse models can offer complementary insights into the data.

In summary, stacking enhances prediction accuracy by combining diverse base models using a meta-learner. Its hierarchical structure, flexibility in model selection, and ability to exploit different inductive biases make it a valuable tool in machine learning. The next section will explore how stacking can be applied in specific domains, highlighting its practical utility and potential for further research.

2.2.3 Uncertainty Quantification

In recent years, the emphasis on building trustworthy and robust deep learning systems has intensified, particularly in domains like biomedical imaging and safety-critical automation. Uncertainty Quantification (UQ) has emerged as a fundamental pillar in addressing this need. Rather than merely relying on point predictions, UQ provides insight into how confident a model is about its outputs, thereby guiding safer and more informed decision-making. This is especially critical in high-risk scenarios such as medical diagnosis, where incorrect predictions made with unwarranted confidence can lead to adverse outcomes.

Uncertainty in machine learning can broadly be categorized into two fundamental types: aleatoric and epistemic uncertainty. Aleatoric uncertainty captures the intrinsic noise or randomness present in the data. This form of uncertainty is irreducible and arises from limitations in the measurement process or inherent variability in

the input domain. On the other hand, epistemic uncertainty represents a model's ignorance or lack of knowledge about the data-generating process. It arises due to insufficient training data or inadequate model capacity and can, in principle, be reduced by acquiring more diverse data or using a more expressive model architecture [26, 37].

A variety of methods have been developed to quantify these two sources of uncertainty. Among these, Bayesian Neural Networks (BNNs) offer a principled probabilistic framework by modeling weights as distributions rather than fixed values. Variational inference techniques and Monte Carlo sampling methods like Hamiltonian Monte Carlo and Langevin dynamics allow the posterior distribution over weights to be approximated efficiently, thereby enabling uncertainty estimation. However, such approaches are often computationally expensive and memory-intensive, particularly when scaling to deep architectures. As an alternative, more tractable solutions such as Monte Carlo Dropout have been proposed. These methods approximate the predictive distribution by performing multiple stochastic forward passes with dropout activated at test time, effectively simulating sampling from a posterior [26].

Another class of approaches relies on conformal prediction (CP), which offers a distribution-free method of producing statistically valid prediction sets without strong assumptions about the underlying model or data distribution. In contrast to Bayesian methods that estimate full predictive distributions, CP produces a set of candidate labels for a given input, guaranteed to contain the true label with a user-defined probability. Recent research has validated the effectiveness of CP in classification problems, particularly for medical imaging tasks such as skin lesion diagnosis, demonstrating that CP can outperform traditional uncertainty estimation methods like Monte Carlo Dropout and evidential learning in terms of reliability and robustness [16].

2.2.3.1 Conformal Prediction

Because of its noteworthy characteristics, CP is a new technique for measuring uncertainty that has attracted a lot of interest in the computer vision and machine

learning fields. The prediction set of all possible labels that includes the true label with a user-defined confidence level is the result of the CP algorithm.

CP techniques have risen to prominence due to their reputation for being easy to implement, computationally efficient, and adaptable across various models [24]. Their applicability extends across a wide array of pre-trained models, encompassing decision support models, shallow multi-layer perceptrons, deep neural networks, and beyond. Moreover, CP methodologies find utility across various tasks such as classification or regression, demonstrating their versatility [4].

An additional benefit is that CP approaches are robust when handling various data kinds and structures because they do not rely on assumptions regarding the underlying data distribution. Moreover, they provide a validity guarantee even in situations when there is a limited quantity of samples, highlighting their dependability in practical settings. As a consequence, CP becomes an essential tool in a wide range of real-world machine learning and computer vision applications [5]. To facilitate the implementation of the CP algorithm, the next section delves into the intricate details.

To delve into the intricacies of CP, let us consider a classification task where a machine learning model is trained on an image dataset, denoted as X_{train}, aiming to accurately classify each input image into one of T total classes, with corresponding labels Y_{train}. The SoftMax function σ, applied to the class probability outputs $f(x)$, serves as an indicator of model uncertainty, generating conformity scores denoted as $S{=}\sigma f(X)$. Lastly, the calibration dataset can be used as an input to the trained model in order to generate the empirical quantile ($\hat{q}$) [16]. With the knowledge of $\hat{q}$, the CP module generates a prediction set that includes the correct class along with its associated confidence level. In this context, for a given test image, the conformity score is calculated, and subsequently, the prediction set is determined as follows.

$$C(X_{\text{test}}) = \{y : S(X_{\text{test}}) < \hat{q}\} \tag{2.1}$$

$S(X_{\text{test}})$ is the Cumulative SoftMax output of a test sample; therefore, the prediction set includes the classes, that the SoftMax output is less than the 1 - $\hat{q}$. In this manner,

the classifier model plays a crucial role in the CP framework. Thus, ensuring the accuracy of the model is paramount for generating a meaningful $\hat{q}$ and an accurate prediction set. Using S and this calibration data, our goal is to generate a prediction set of possible labels $C(X_{\text{test}}) \subset \{1, \ldots, K\}$ that is valid in the following sense [6]:

$$[1 - \alpha \leq P(Y_{\text{test}} \in C(X_{\text{test}})) \leq 1 - \alpha + \frac{1}{n+1} \tag{2.2}$$

where $(X_{\text{test}}, Y_{\text{test}})$ is a new test point from the same distribution, and $\alpha \in [0, 1]$ is a user-defined error rate.

2.2.3.2 Adaptive Prediction Sets (APS)

To address the limitations of CP in scenarios with heterogeneous prediction difficulty, the Adaptive Prediction Sets (APS) method was introduced. APS adapts the size of the prediction set to the level of uncertainty associated with each test input. Rather than using a fixed threshold across all samples, APS allows the prediction set to dynamically expand or contract based on the individual softmax distribution, offering more informative uncertainty estimates.

In APS, the conformity score is defined as the cumulative probability mass of all classes with probabilities equal to or higher than the true class:

$$S_{\text{APS}}(x, y) = \sum_{j: p_j \geq p_y} p_j \tag{2.3}$$

This score measures how dominant the true class is within the softmax distribution. During calibration, these scores are collected over the validation set, and the empirical quantile $\hat{q}_{\text{APS}}$ is computed. For a test sample, all classes with cumulative probabilities below this threshold are included in the prediction set:

$$C_{\text{APS}}(x) = \left\{ y : \sum_{j: p_j \geq p_y} p_j \leq \hat{q}_{\text{APS}} \right\} \tag{2.4}$$

APS tends to produce tighter prediction sets for confident samples and larger

sets for ambiguous ones, thereby improving the efficiency and informativeness of the predictions. However, it remains susceptible to the same calibration issues as standard softmax classifiers. If the softmax outputs are poorly calibrated, the cumulative score may not reflect true uncertainty, leading to either overconfident or overly large prediction sets [16].

2.2.3.3 Regularized Adaptive Prediction Sets (RAPS)

Regularized Adaptive Prediction Sets (RAPS) refine the APS method by introducing a regularization term to penalize the size of the prediction set explicitly. This modification aims to reduce the frequency of large, uninformative prediction sets without compromising the theoretical coverage guarantees.

The modified conformity score in RAPS takes the form:

$$S_{\text{RAPS}}(x, y) = \sum_{j: p_j \geq p_y} p_j + \lambda \cdot |C(x)| \tag{2.5}$$

Here, λ is a non-negative regularization parameter that controls the trade-off between validity and set compactness. By adding this term, RAPS encourages the algorithm to prefer smaller prediction sets unless the conformity score strongly justifies a larger set. This leads to a more efficient use of uncertainty information, particularly in multi-class classification problems where APS may otherwise default to selecting many classes.

Despite its advantages, RAPS introduces an additional hyperparameter λ that requires tuning. Over-regularization can lead to under-coverage, especially in uncertain scenarios, while under-regularization reduces the benefits of the method. Thus, RAPS achieves the best results when used in conjunction with well-calibrated models and a careful validation of hyperparameters [16].

2.2.3.4 Monte Carlo Dropout (MCD)

Monte Carlo Dropout (MCD) provides a scalable approximation to Bayesian inference by employing dropout at test time to simulate sampling from the posterior

distribution over model parameters. Originally proposed by Gal and Ghahramani, this method leverages the equivalence between dropout and variational inference to capture model uncertainty without requiring explicit posterior distributions.

In practice, MCD involves running the same input through the model multiple times with different dropout masks. This produces a set of softmax predictions $\hat{p}_1, \ldots, \hat{p}_T$, which are then aggregated to compute the mean and variance:

$$\bar{p} = \frac{1}{T} \sum_{t=1}^{T} \hat{p}_t, \quad \text{Var}(\hat{p}) = \frac{1}{T} \sum_{t=1}^{T} (\hat{p}_t - \bar{p})^2 \tag{2.6}$$

The mean prediction $\bar{p}$ is used for classification, while the variance estimates the model's epistemic uncertainty. MCD is attractive for its simplicity and compatibility with standard deep learning pipelines, as it does not require retraining or architectural changes [44].

Nevertheless, MCD has limitations. Its computational cost increases linearly with the number of forward passes T, which can hinder real-time applications. Moreover, it only captures epistemic uncertainty and fails to model aleatoric uncertainty, which is crucial in noisy environments. Lastly, MCD assumes that dropout provides a good variational approximation, which may not always hold, especially for complex tasks [16].

2.2.3.5 Softmax Thresholding

A simpler yet less rigorous method for UQ is softmax thresholding, wherein the model predicts a label only if the maximum softmax score exceeds a predefined threshold τ. Otherwise, the prediction is flagged as uncertain or abstained. This method is straightforward and does not require a calibration set or multiple passes.

Formally, the predicted class y^* is accepted only if:

$$\max_k p_k(x) > \tau \tag{2.7}$$

Softmax thresholding is commonly used in practical systems due to its ease of implementation and low computational overhead. However, it lacks formal guaran-

tees on coverage or calibration. Neural networks are known to be poorly calibrated, especially when trained using standard cross-entropy loss, often leading to overconfident predictions even on incorrect inputs. Consequently, thresholding may yield misleading uncertainty estimates in practice [16].

2.2.4 Vision Transformers for Skin Lesion Classification

Vision Transformers (ViTs) have emerged as a transformative model architecture in the field of computer vision, offering a compelling alternative to Convolutional Neural Networks (CNNs). Their unique ability to capture global contextual information using self-attention mechanisms makes them particularly suitable for medical imaging tasks such as skin lesion classification, where capturing subtle and spatially distant visual cues is critical. Traditional CNNs, despite their success in visual tasks, are inherently limited by their local receptive fields and fixed inductive biases. These constraints restrict their ability to model long-range dependencies across an image, which is essential in differentiating nuanced features between benign and malignant skin lesions. In contrast, ViTs mitigate these limitations by treating an image as a sequence of patches, enabling a holistic analysis of the visual data [14, 21].

To process visual data, ViTs begin by partitioning an input image $x \in \mathbb{R}^{H \times W \times C}$ into a grid of non-overlapping patches of size $P \times P$. Each patch is flattened into a one-dimensional vector of length $P^2 \cdot C$ and then linearly projected into a D-dimensional embedding space using a trainable weight matrix $E \in \mathbb{R}^{(P^2 \cdot C) \times D}$. This transformation creates a sequence of $N = \frac{HW}{P^2}$ patch embeddings. A special classification token [CLS] is prepended to the sequence, and a learnable positional embedding matrix $E_{pos} \in \mathbb{R}^{(N+1) \times D}$ is added to encode positional information that would otherwise be lost due to the permutation-invariant nature of self-attention:

$$Z_0 = [x_{\text{CLS}}; z_p^1; z_p^2; \ldots; z_p^N] + E_{pos}$$

This formulation allows the model to preserve spatial relationships while leveraging the power of sequence modeling.

The input sequence Z_0 is fed into a series of L Transformer encoder blocks. Each

block comprises two primary sublayers: a Multi-Head Self-Attention (MSA) mechanism and a position-wise Feed-Forward Network (FFN). Each sublayer is followed by residual connections and layer normalization. The self-attention operation allows each patch to interact with all others, enabling the model to learn dependencies and contextual information over the entire image:

$$Z'_l = \text{MSA}(\text{LN}(Z_{l-1})) + Z_{l-1}$$
$$Z_l = \text{FFN}(\text{LN}(Z'_l)) + Z'_l$$

The self-attention is defined mathematically as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where Q, K, and V are query, key, and value matrices obtained by projecting the input embeddings. The division by $\sqrt{d_k}$ stabilizes the gradients and the softmax ensures a normalized attention distribution.

To enhance the model's expressivity, ViTs utilize multi-head attention, where the attention mechanism is computed h times in parallel with different parameter sets. Each head operates in a different subspace of the input representation and captures diverse aspects of the image context. The outputs of these attention heads are then concatenated and linearly projected to match the input dimension:

$$\text{MSA}(X) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O$$

This design allows the model to jointly attend to information from different representation subspaces, improving its capacity to model complex visual relationships.

The final embedding corresponding to the [CLS] token, denoted as Z_L^{CLS}, is used as a holistic representation of the input image. This vector is passed through a multilayer perceptron (MLP) head to generate the class probabilities. The classification head usually comprises one or more fully connected layers followed by a softmax

activation:

$$\hat{y} = \text{softmax}(W_{\text{cls}} Z_L^{\text{CLS}} + b)$$

This approach enables the ViT to produce predictions based on a globally aggregated feature representation, in contrast to CNNs, which typically rely on local features aggregated via pooling operations.

In the domain of medical image analysis, ViTs offer several compelling advantages. The capacity to capture long-range dependencies makes them well-suited for tasks like skin lesion classification, where malignant features may be subtle and spatially distributed. Moreover, ViTs are inherently more interpretable, as attention maps can be visualized to indicate which parts of the image contributed most to the model's decision. This transparency is crucial in clinical settings, where understanding the rationale behind a diagnosis can inform treatment decisions and foster trust in AI systems [31, 22].

Furthermore, the modular nature of ViTs allows for flexible architectural modifications. Researchers can adjust the patch size, model depth, and attention heads to balance accuracy and computational efficiency. When pretrained on large-scale datasets and fine-tuned on specialized medical datasets, ViTs can achieve performance that rivals or exceeds that of CNN-based models, even when training data is scarce [40].

To overcome the data-hungriness and computational demands of the original ViT, several architectural variants have been proposed. For instance, Data-efficient Image Transformers (DeiT) incorporate a distillation token during training, enabling the model to learn effectively from smaller datasets. Swin Transformers introduce a hierarchical structure with shifted window attention to capture both local and global features while maintaining manageable computational complexity. Hybrid architectures that combine CNN-based feature extractors with transformer encoders have also demonstrated improved generalization by leveraging the strengths of both paradigms [21].

These innovations have significantly broadened the applicability of ViTs, making them more accessible for medical imaging tasks. For example, Swin Transformers

can adapt to various input resolutions, a desirable feature when dealing with high-resolution dermoscopic images. DeiT, with its lightweight configuration, allows researchers to deploy transformer-based models in resource-constrained environments such as mobile diagnostic tools.

Despite their promise, ViTs are not without limitations. Their reliance on large-scale pretraining poses challenges in domains where such datasets are unavailable or domain-specific. Additionally, the lack of inherent inductive biases such as locality and translation invariance makes ViTs less efficient than CNNs when trained from scratch. Another critical concern is the high computational cost, especially in early layers where the number of tokens is large.

Future research may focus on developing more efficient ViT architectures tailored for medical imaging. This includes integrating domain-specific knowledge, leveraging self-supervised learning, and designing adaptive token pruning strategies. Such enhancements could reduce computational overhead while preserving or improving classification performance. Moreover, combining ViTs with uncertainty estimation techniques—such as conformal prediction—can make the models not only accurate but also reliable and trustworthy in clinical applications.

In conclusion, Vision Transformers represent a significant step forward in the field of medical image classification. Their ability to integrate global image context, combined with their flexibility and interpretability, makes them a valuable tool for developing robust AI systems in healthcare. As the field evolves, continued innovations in model architecture, training strategies, and domain adaptation will further enhance their utility in skin lesion analysis and beyond.

2.3 Contributions of This Thesis

The contributions of this thesis are aligned with addressing the critical gaps in skin lesion classification and advancing the state of the art in medical image analysis:

1. **Framework Development:** Introduction of **EACE-ViTs**, a novel framework that combines Vision Transformers, entropy-based ensemble weighting,

and Generative Adversarial Networks (GANs) for robust data augmentation.

2. **Uncertainty Quantification:** Development of dynamic entropy-adjusted thresholds for conformal prediction methods, enabling calibrated and adaptive prediction sets.

3. **Comprehensive Evaluation:** Extensive validation of the framework on the HAM10000 dataset, with comparisons to state-of-the-art methods in accuracy, coverage, and interpretability metrics.

4. **Impactful Insights:** Demonstration of the clinical applicability of the framework, emphasizing its potential to improve diagnostic confidence and decision-making in dermatology.

Chapter 3

Entropy-Aware Conformal Ensemble of Vision Transformers for Reliable Skin Lesion Classification

3.1 Dataset

This study leverages the HAM10000 dataset [23], a comprehensive dermatoscopic image collection that includes 10,015 images categorized into seven classes of skin lesions. These categories include both benign and malignant lesions, reflecting the diversity encountered in real-world dermatology practice. The dataset serves as a benchmark for automated skin lesion classification and poses several challenges due to its visual variability, class imbalance, and clinical overlap between categories. To ensure robust evaluation, we divided the dataset into four distinct subsets: 50% for training, 20% for validation, 20% for calibration, and 10% for testing. All test results reported in this thesis are based solely on the test partition to ensure unbiased performance estimation. In what follows, we describe each class in detail, accompanied by a representative dermatoscopic image from the HAM10000 dataset [10, 36, 12, 28, 30] .

3.1.1 Actinic Keratoses and Intraepithelial Carcinoma (akiec)

Figure 3.1: Actinic keratoses and intraepithelial carcinoma (akiec)

Actinic keratoses (AK) are precancerous skin lesions caused primarily by long-term exposure to ultraviolet (UV) radiation. They often present as rough, scaly patches on sun-damaged skin, particularly on the face, ears, and forearms. In some cases, AK can progress to squamous cell carcinoma in situ, known as intraepithelial carcinoma (IEC). Dermoscopically, these lesions show erythematous backgrounds with scale and irregular vascular structures. The visual overlap with other keratotic lesions and early squamous cell carcinoma makes diagnosis challenging. Treatment typically involves cryotherapy, topical agents like 5-fluorouracil, or photodynamic therapy. Due to their intermediate position between benign and malignant conditions, accurate identification of akiec lesions is critical for timely intervention.

3.1.2 Basal Cell Carcinoma (bcc)

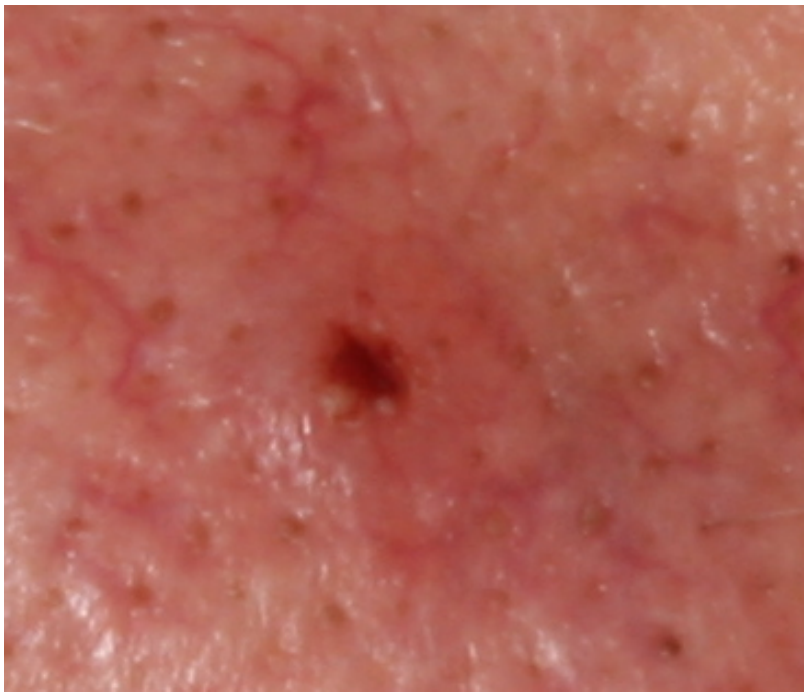


Figure 3.2: Basal cell carcinoma (bcc)

Basal cell carcinoma is the most common form of skin cancer, often arising from chronic sun exposure in fair-skinned individuals. Though it rarely metastasizes, BCC can be locally invasive and cause significant tissue damage if left untreated. It frequently appears as a pearly papule with rolled borders, telangiectasia, or ulceration. In dermoscopic images, BCC often shows arborizing vessels, leaf-like areas, or spoke-wheel structures. Its visual resemblance to benign lesions, such as seborrheic keratoses or nevi, complicates early diagnosis. Standard treatments include surgical excision, Mohs micrographic surgery, and topical agents like imiquimod for superficial types. Accurate detection of BCC is essential to avoid disfigurement and recurrence.

3.1.3 Benign Keratosis-like Lesions (bkl)

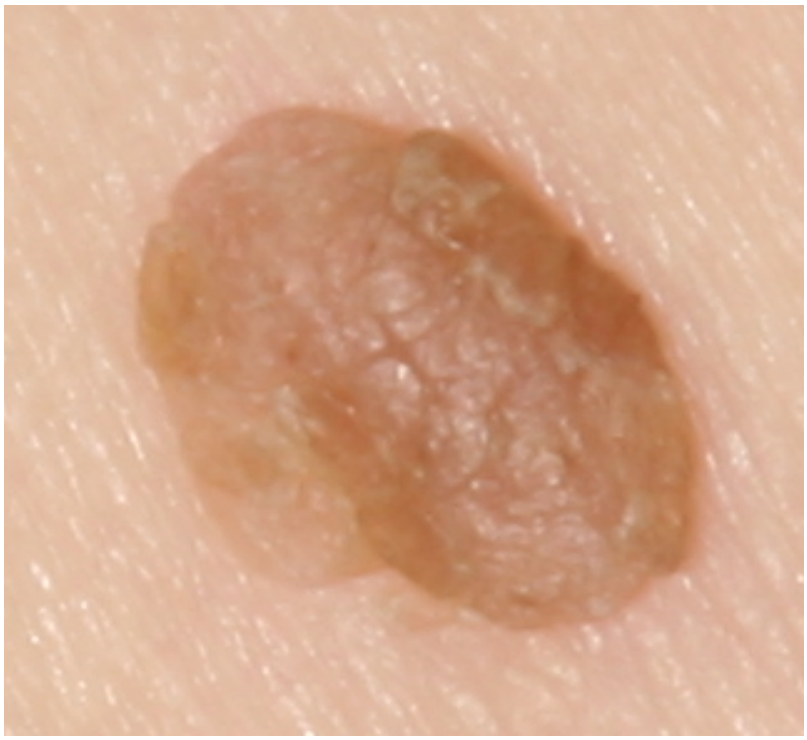


Figure 3.3: Benign keratosis-like lesion (bkl)

Benign keratosis-like lesions encompass several subtypes, including seborrheic keratoses, solar lentigines, and lichen planus-like keratoses. These lesions are non-cancerous and often appear as warty, brown plaques with well-demarcated borders. Dermoscopically, they can exhibit milia-like cysts, comedo-like openings, or cerebriform patterns. Despite being benign, bkl lesions are frequently misclassified as melanomas or basal cell carcinomas due to their color variability and texture. As such, they pose a significant challenge for both human and machine-based classification systems. While they generally require no treatment, accurate diagnosis is important to avoid unnecessary biopsies or surgical procedures.

3.1.4 Dermatofibroma (df)

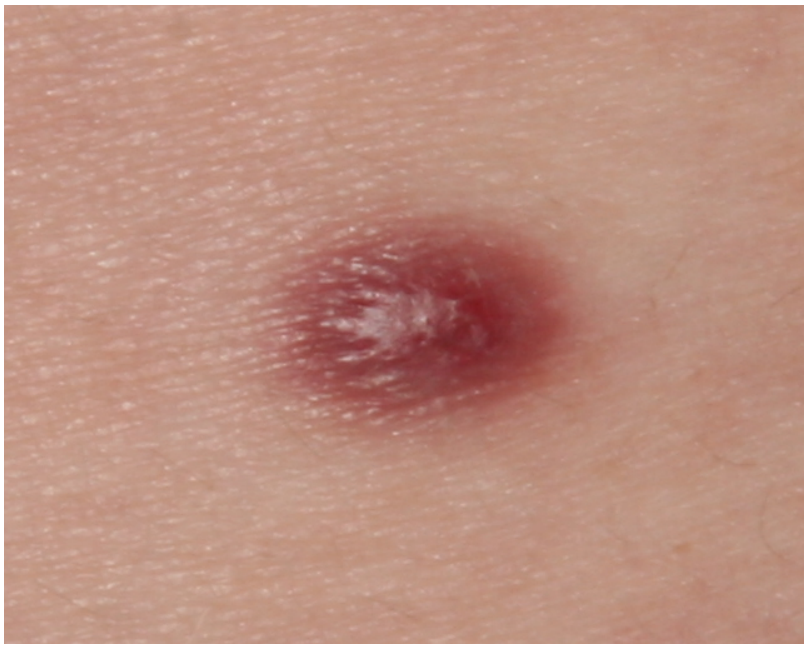


Figure 3.4: Dermatofibroma (df)

Dermatofibroma is a benign fibrohistiocytic skin tumor that typically occurs on the lower extremities of adults. Clinically, it presents as a firm, hyperpigmented nodule, often with a central dimple when compressed. Under dermoscopy, dermatofibromas often display a central white scar-like area surrounded by a delicate pigment network. Although benign, their appearance can resemble melanomas or other pigmented lesions, especially when atypical patterns are present. Due to their rarity in most image datasets, they are typically underrepresented, which makes training classifiers to detect them effectively a challenge. Treatment is not usually necessary unless for cosmetic reasons or diagnostic uncertainty.

3.1.5 Melanoma (mel)

Figure 3.5: Melanoma (mel)

Melanoma is an aggressive and potentially lethal form of skin cancer that arises from melanocytes. It is responsible for the majority of skin cancer-related deaths worldwide. The lesion may exhibit asymmetry, irregular borders, and multiple colors, ranging from black and brown to red, blue, or white. Dermoscopically, melanoma can present with atypical pigment networks, blue-white veils, irregular dots and globules, and regression structures. Early detection is critical, as melanoma is highly curable when diagnosed at an early stage but becomes life-threatening once metastasis occurs. Treatments include surgical excision, immunotherapy, and targeted therapies such as BRAF inhibitors. However, its visual similarity to benign nevi, particularly in early stages, poses one of the most difficult challenges in dermatology.

3.1.6 Melanocytic Nevus (nv)

Figure 3.6: Melanocytic nevus (nv)

Melanocytic nevi, commonly known as moles, are benign proliferations of melanocytes. They are often symmetric, uniformly pigmented, and well-circumscribed. Despite their benign nature, nevi are the most commonly confused lesion with melanoma,

particularly when they exhibit atypical features such as asymmetry or irregular pigmentation. Dermoscopic patterns may include reticular networks, globular structures, or homogeneous pigmentation. Because nevi dominate the HAM10000 dataset, they often introduce class imbalance in training, which can bias classifiers to overpredict this category. Nonetheless, accurate classification is important, especially when evaluating borderline lesions or monitoring changes over time.

3.1.7 Vascular Lesions (vasc)

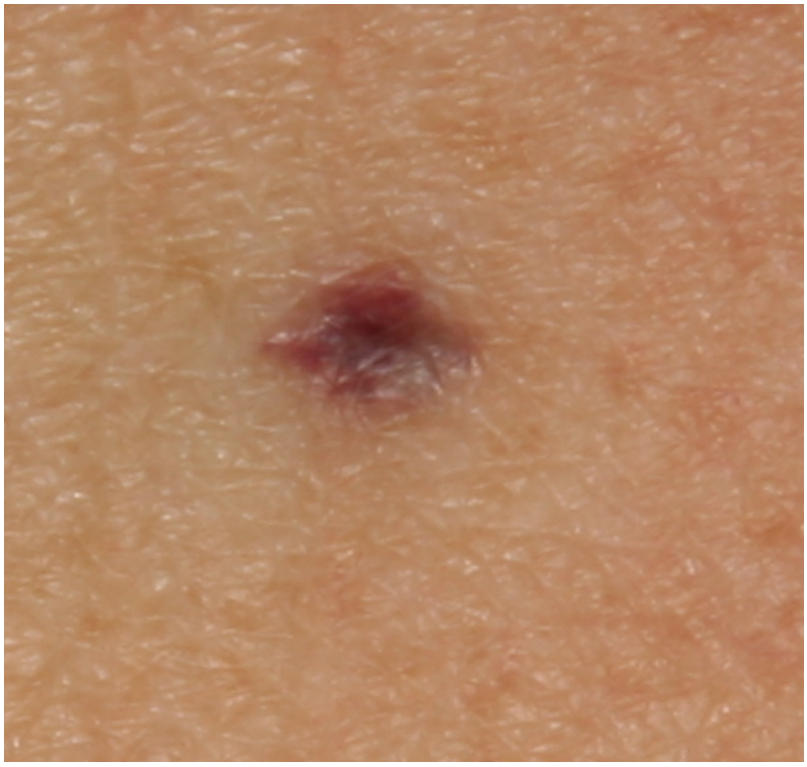


Figure 3.7: Vascular lesion (vasc)

Vascular lesions in the HAM10000 dataset include angiomas, hemangiomas, and angiokeratomas. These lesions typically present as bright red, purple, or blue papules or macules, depending on the depth and type of vascular involvement. Dermoscopic features include red lacunae, red-blue structureless areas, or serpentine vessels. Though

often benign, they can be mistaken for pigmented lesions or nodular melanomas due to overlapping visual patterns. Treatments for vascular lesions vary based on type and severity, ranging from observation to laser therapy or surgical removal. Their relatively low occurrence in training data makes them a difficult class for automated classifiers to learn robustly.

3.2 GAN-Based Data Augmentation

To address the severe class imbalance in the HAM10000 dataset, we employed a Generative Adversarial Network (GAN) to synthesize new samples for underrepresented lesion categories. The class distribution prior to augmentation was highly skewed, with a majority of the dataset dominated by a single class (e.g., `nv`), while minority classes such as `df`, `vasc`, and `akiec` contained significantly fewer instances. This imbalance is problematic for deep learning models, especially in medical image classification, where fair representation is critical for achieving robust and unbiased predictions across all lesion types.

Our GAN model comprised a Generator (G) and a Discriminator (D), trained in a min-max adversarial setting. The objective of the Generator was to create realistic synthetic images from noise and transformations of existing lesion samples, while the Discriminator attempted to differentiate between real and generated images. The adversarial training was governed by the following loss functions:

$$\mathcal{L}_D = -\mathbb{E}_{x\sim p_{\text{data}}}[\log D(x)] - \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))] \tag{3.1}$$

$$\mathcal{L}_G = -\mathbb{E}_{z\sim p_z}[\log D(G(z))] \tag{3.2}$$

The loss functions presented above describe the adversarial objectives for the Generator and Discriminator within the GAN framework. In these expressions, $x \sim p_{\text{data}}$ refers to a real image sampled from the true data distribution, while $z \sim p_z$ denotes a noise vector or a transformed sample input to the Generator. The Generator, denoted by $G(z)$, takes this input and produces a synthetic image. The

Discriminator, represented by $D(\cdot)$, outputs a probability indicating whether the given image is real or generated. Specifically, $D(x)$ is the Discriminator's estimated probability that a real image x is authentic, and $D(G(z))$ is its estimate that the synthetic image $G(z)$ is real. The Discriminator's loss function $\mathcal{L}_D$ encourages $D(x)$ to be close to 1 (correctly identifying real images) and $D(G(z))$ to be close to 0 (correctly rejecting fake images). In contrast, the Generator's loss $\mathcal{L}_G$ drives $D(G(z))$ toward 1, attempting to deceive the Discriminator into classifying synthetic images as real. This adversarial training setup fosters a continuous competition in which the Discriminator becomes more adept at distinguishing real from fake, and the Generator learns to produce increasingly realistic synthetic lesions. This dynamic is fundamental to generating high-fidelity synthetic images that effectively enhance the diversity and balance of the training dataset.

We used Binary Cross-Entropy (BCE) loss for both networks, optimized using the Adam optimizer with a learning rate of 0.0002 and $\beta_1 = 0.5$. The GAN was trained for 50 epochs exclusively on images from underrepresented classes, ensuring that the Generator learned a rich distribution without overfitting to scarce training samples.

Upon completion of training, the Generator produced up to 300 high-resolution synthetic images per minority class. These samples underwent visual inspection to ensure clinical plausibility and diversity. The final filtered synthetic images were appended to the original training set, thereby enhancing both class balance and data variability.

To evaluate the effect of GAN-based augmentation on dataset distribution, we generated pie charts illustrating the class proportions before and after synthetic image integration. Figure 3.8 shows the relative sizes of each class in the training set. As evident from the left chart, classes like `nv` dominate the original dataset with over 60% of the total samples, while classes like `df` and `vasc` occupy less than 2% each.

The right chart illustrates the post-augmentation class distribution. While we aimed for a near-balanced distribution, we intentionally allowed for slight variation across categories (5–15%) to retain a degree of realism and avoid potential overfitting

that may arise from fully synthetic balancing. This soft balancing simulates natural variance while improving fairness in model training.

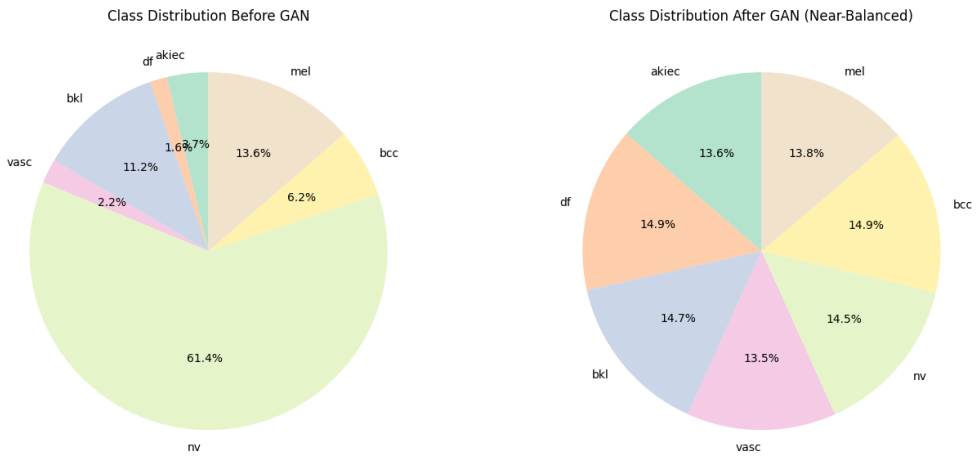


Figure 3.8: Class distribution in the HAM10000 dataset before (left) and after (right) GAN-based augmentation. The dataset initially exhibits strong imbalance, which is mitigated by adding synthetic images to underrepresented classes, resulting in a more uniform and realistic distribution.

This augmentation strategy significantly improved the performance of the EACE-ViTs (Entropy-Aware Conformal Ensemble of Vision Transformers) framework, particularly for classes previously lacking sufficient training examples. Empirically, the ensemble showed increased per-class accuracy and better-calibrated confidence intervals, with minority classes such as `akiec` and `bkl` benefiting the most from the added diversity. The improved class distribution also reduced the prediction entropy across these classes, making the conformal prediction intervals more informative and compact.

Furthermore, balanced representation allowed for more effective ensemble agreement, reducing the reliance on dominant-class voting. The softmax entropy filtering module, when applied to this enriched dataset, showed higher confidence in its selections, as evident in lower calibration error rates in subsequent experiments.

In summary, our GAN-based augmentation technique effectively mitigated class

imbalance while preserving visual and clinical fidelity. By generating realistic and diverse lesion images, we not only balanced the dataset but also enhanced the robustness and generalizability of our transformer-based classification framework. The impact of this step is crucial in clinical settings, where equitable performance across all lesion types is essential for trustworthy automated diagnostics.

3.3 Entropy-Aware Conformal Ensemble of Vision Transformers (EACE-ViTs) Framework

In this paper, we present an entropy-aware conformal ensemble learning system, EACE-ViTs, which is rigorously developed to achieve robust and reliable classification of skin lesions. The system comprises several unique components that act in concert to boost overall performance: first, data augmentation via GAN is employed to effectively alleviate data imbalances in the HAM10000 dataset to achieve proportional representation of all lesion classes. Subsequently, the framework employs a number of ViT models to conduct precise feature extraction, leveraging their capacity for identifying faint patterns and textures in medical images. To enhance decision-making further, entropy-based ensemble weighting is employed, where weights are updated dynamically based on the confidence values of model predictions. Finally, the framework integrates entropy-adjusted thresholds into two state-of-the-art CP approaches—RAPS and APS—and thereby enhances the reliability of uncertainty quantification. Each step is meticulously crafted to build upon the predecessor such that enhancements to the data enhance the quality of the inputs to the model, thereby enhancing the strength and validity of the predictions.

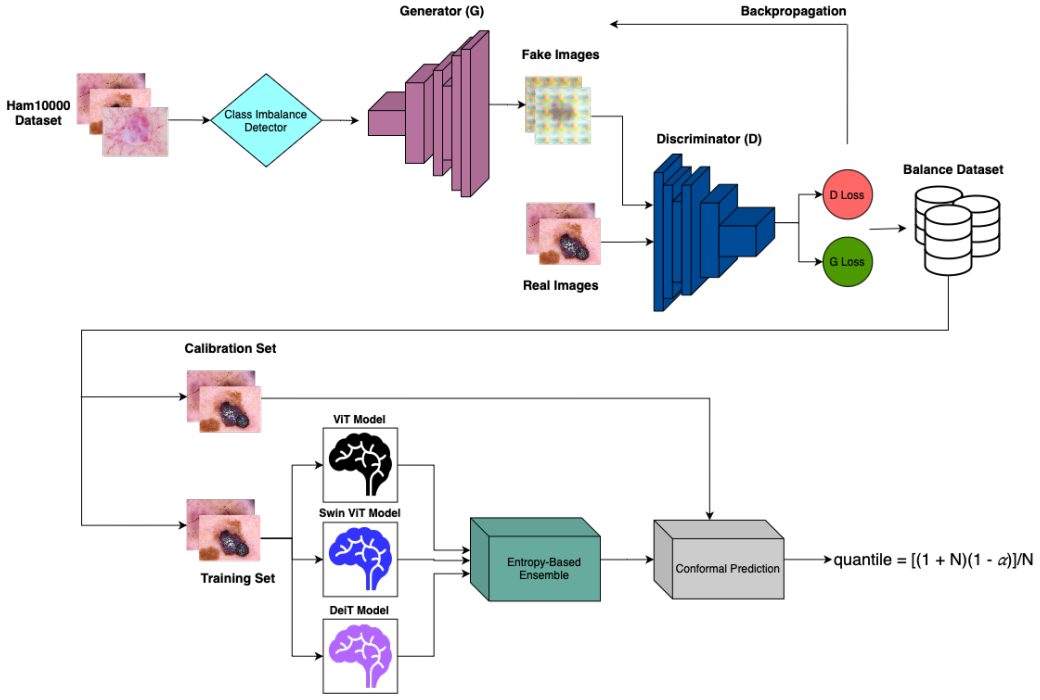


Figure 3.9: Training stage of the EACE-ViTs framework.

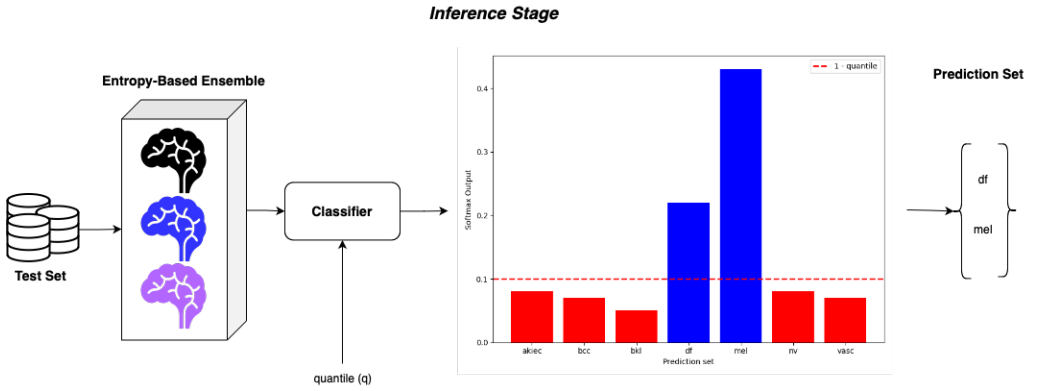


Figure 3.10: Inference stage of the EACE-ViTs framework.

EACE-ViTs begins with class imbalance detection in the HAM10000 dataset. The classes identified as underrepresented are amplified through a GAN model, which generates synthetic samples to construct class-balanced distributions. This step enhances the model's generalizability across all lesion types, a common issue

with medical imaging datasets.

After preparing the balanced dataset, we move on to model training. A collection of Vision Transformer models with different architectures is employed, i.e., *vit-base-patch16-224*, *swin-base-patch4-window7-224*, and *deit-base-patch16-224*, with each model trained individually on the balanced dataset. These models obtained from the timm library are chosen given their strong capacity in handling complicated image data, thereby guaranteeing extensive feature extraction in various dimensions of skin lesion images. This design choice allows for leveraging the individual advantages of each transformer structure to improve the diagnostic accuracy and reliability of our ensemble model. The models capture diverse features of skin lesion characteristics, thus improving the overall robustness of the framework.

An entropy-based method is employed for the ensemble. Each model calculates the softmax probabilities for each input, and the entropy of the probabilities is used as a measure of confidence of each model's predictions. The model weights in our ensemble, i.e., 0.8 for the least entropy model and 0.1 for the other models, were established using an empirical strategy that favored a trial-and-error process. This process entailed extensive experimentation with different weighting to ascertain the optimal distribution that enhances the collective predictive accuracy and reliability of the ensemble. We performed an analysis of different weight combinations via several validation iterations, observing their impact on the performance measures of the ensemble, such as accuracy, precision, recall, and F1-score, against the specific challenges of skin lesion classification. The weights of 0.8 and 0.1 used here have consistently shown improved performance, balancing the influence of the most confident model while considering the influence of the other models in the ensemble. This strategy allows our ensemble to reap the benefits of the strengths of the most confident predictions without depending too much on any one model, thereby increasing robustness and preventing overfitting of specific characteristics or biases in the training data. The ensemble's final prediction is then a weighted average of the

component models' predictions:

$$p_{\text{ensemble}}(y|x) = \sum_{m=1}^{3} w_m p_m(y|x)$$

where $w_{m^*} = 0.8$ for the model with the lowest entropy m^*, and $w_{j \neq m^*} = 0.1$ for the others. This entropy-aware weighting allows the framework to prioritize predictions with higher confidence, thus enhancing reliability.

The last step consists of Conformal Prediction (CP), which is employed to assess the uncertainty present in the ensemble predictions. The data is divided into four subsets: training, validation, calibration, and test sets, where 50% is allocated for training, 20% for validation, 20% for calibration, and 10% for testing. Nonconformity scores of predictions from all models are computed based on the calibration set. For reliability enhancement, the CP model uses entropy-normalized thresholds and thereby enhances prediction intervals for accuracy. CP is then utilized to generate calibrated sets of predictions, which guarantees that every set contains the true label at a given level of confidence. For this purpose, the Regularized Adaptive Prediction Sets (RAPS) and Adaptive Prediction Sets (APS) approach is used, which involves entropy-based adaptations to make the prediction models more resilient and adaptive.

The complete framework, encompassing all the stages, is presented in Algorithm (1) and is also pictorially represented in Fig. 3.9 and Fig. 3.10. As evident in Fig. 3.9 and Fig. 3.10, the EACE-ViTs framework suggests a GAN-based augmentation policy, entropy-based ensemble weighting, and CP-based uncertainty quantification. Through the use of entropy-aware ensemble weighting and conformal prediction in conjunction with Vision Transformers, EACE-ViTs formulates a solid, uncertainty-calibrated system that is well-suited for high-stakes environments such as medical diagnosis. The system not only improves predictive performance but also gives realistic estimates of uncertainty, thereby rendering it an innovative method of reliable classification of skin lesions in high-stakes healthcare applications.

Algorithm 1 Proposed Algorithm: Entropy-Aware Conformal Ensemble of Vision Transformers (EACE-ViTs)

Input: Dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, class imbalance threshold τ, models $\mathcal{M} = \{$ViT, Swin, DeiT$\}$, scaling factor λ, index of the class labels k

Output: Prediction sets with APS and RAPS

Step 1: Addressing Class Imbalance class c in $\mathcal{D}$ $|c| < \tau$ Augment $\mathcal{D}$ with GAN-generated samples for class c

Step 2: Model Training model $m \in \mathcal{M}$ Train m on $\mathcal{D}$ to obtain predictions $f_m(x)$

Step 3: Entropy-Based Ensemble each $x \in \mathcal{D}$

Compute $\{H_m\}_{m \in \mathcal{M}}$, where $H_m = -\sum_k p_m(k|x) \log p_m(k|x)$

Assign weights: $w_{m^*} = 0.8$ for $m^* = \arg\min H_m$, $w_{j \neq m^*} = 0.1$

Ensemble prediction:

$$p_{\text{ensemble}}(y|x) = \sum_{m \in \mathcal{M}} w_m p_m(y|x)$$

Step 4: Entropy-Adjusted Conformal Prediction each calibration sample $(x, y) \in \mathcal{C}$

Compute entropy $H(x)$ and normalize using the maximum possible entropy: $H_{\text{norm}} = \frac{H(x)}{H_{\max}}$

Adjust threshold: $\tau_{\text{adjusted}} = \tau + \lambda \cdot H_{\text{norm}}$

Generate prediction sets using APS and RAPS methods

3.4 Configuration and Training Setup

In this study, three pre-trained models—ViT, Swin Transformer, and DeiT—were utilized, sourced from the Timm library in Python [41]. Leveraging transfer learning with these pre-trained models enables effective adaptation to the new dataset. To facilitate this, the weights of the backbone layers were frozen, while the MLP head was replaced with a new classifier head, allowing the models to retain valuable learned

features while tailoring the classification layers to the specific characteristics of the skin lesion dataset.

During training, the Adaptive Moment Estimation (ADA) optimizer was used with a learning rate set to 0.0001. The models were trained over 5 epochs, with cross-entropy as the loss function. This setup aims to balance effective learning with stability, ensuring that each model adapts well to the new task without overfitting.

The computational experiments were conducted on a high-performance system equipped with an AMD Ryzen 9 3950X 16-Core processor and an NVIDIA Titan RTX GPU with 24 GB of dedicated memory. This setup provided the necessary multi-threading capabilities and computational power to handle the intensive training tasks, ensuring consistency and reliability in the experimental results. All experiments were executed on this system, maintaining uniformity throughout the study.

Chapter 4

Results and Discussion

Table 4.1 highlights the performance of the proposed EACE-ViTs framework compared with existing cutting-edge machine learning models for skin cancer classification. Esteva et al.'s early work [15] laid the foundation for automated skin cancer classification using CNNs, achieving 72.10% accuracy for three classes and 55.40% accuracy for nine classes. However, as these models did not incorporate any form of uncertainty estimation, they were limited in providing clinicians with the confidence level necessary for high-stakes decision-making, particularly as the number of classes increased.

Several recent methods have focused on uncertainty quantification to improve the trustworthiness of predictions. For example, Mobiny et al. [32] utilized a Bayesian DenseNet-169 model that not only achieved an accuracy of 83.59% but also introduced uncertainty estimation to enhance clinical interpretability. This approach marked a significant improvement, as it provided both prediction accuracy and insight into model confidence, an essential feature in medical applications.

Bologna and Fossati [8] combined CNNs with a Virtual Discretized Interpretable Multi-Layer Perceptron (VDIMLP) to achieve an accuracy of 84.90%, aiming to increase interpretability without addressing uncertainty quantification. Similarly, Combalia et al. [11] used Test Augmentation and Monte Carlo Dropout (TA + MCD) for uncertainty estimation, although their work did not report a specific accuracy metric, making it challenging to directly assess its predictive performance relative to

others.

Among the advanced methods, Pacheco and Krohling's ResNet-50 [35] reached 91.30% accuracy but lacked uncertainty estimation, which, while impressive, limits its application in settings where prediction reliability is crucial. In contrast, the three-way decision-based Bayesian deep learning model (TWDBDL) introduced by Abdar et al. [1] incorporated uncertainty estimation and achieved 88.95% and 90.96% accuracy on two separate datasets, illustrating how uncertainty quantification can support reliable decisions in dermatology.

The EACE-ViTs framework represents a novel advancement in skin lesion classification, achieving the highest accuracy in this comparison at 94.25%. Unlike prior approaches, EACE-ViTs not only provides precise classification but also includes a novel entropy-aware prediction mechanism that dynamically adjusts based on uncertainty, creating a tailored prediction set according to each model's confidence level. This unique capability positions EACE-ViTs as an advanced framework for clinical applications, addressing both the need for high diagnostic accuracy and the requirement for reliable, calibrated uncertainty information. By delivering both precision and dependable confidence intervals, EACE-ViTs meets an essential demand in medical AI, setting a new standard for trustworthy skin lesion classification in critical healthcare environments.

Table 4.2 presents the performance metrics, including accuracy, macro precision, macro recall, and macro F1 scores, for the DeiT, Swin, ViT models, and our proposed EACE-ViTs framework. Notably, the proposed work outperforms the other models across all metrics, achieving the highest accuracy at 94.25%. This suggests a substantial improvement in classification reliability, indicating the effectiveness of the EACE-ViTs framework for skin lesion diagnosis.

The Swin model exhibits strong performance among the individual transformer-based models, achieving an accuracy of 90.06% and macro F1 score of 0.8635. These values suggest that Swin's architectural design effectively captures complex skin lesion features, outperforming the DeiT and ViT models in this task. However, the relatively lower macro recall for Swin compared to our proposed model (0.8537 vs. 0.8991) indicates that while Swin performs well on a balanced set of classes, it may

Table 4.1: Comparison of the performance of the proposed method with some existing ML models used to classify skin cancer data.

Ref	**# of Classes**	**Model**	**Accuracy (%)**	**Uncertainty**
[15]	3	CNN	72.10	No
[15]	9	CNN	55.40	No
[32]	7	Bayesian DenseNet-169	83.59	Yes
[8]	2	CNN + VDIMLP[1]	84.90	No
[11]	9	TA + MCD[2]	N/A	Yes
[27]	2	CNN	82.90	No
[35]	9	ResNet-50	91.30	No
[34]	2	CNN	60.00	No
[7]	2	SVM[3]	86.00	No
[25]	7	CNN	86.50	No
[39]	7	STCN[4]	80.60	No
[1]	2	TWDBDL	90.96	Yes
Our Work	7	**EACE-ViTs**	**94.25**	**Yes**

miss certain minority class lesions, a critical issue in medical diagnostics.

DeiT performs consistently across metrics, with an accuracy of 89.07% and a macro F1 score of 0.8347. Its macro precision and recall are balanced at 0.8569 and 0.8199, respectively, suggesting that it provides reliable classifications, though it falls short in both precision and recall compared to Swin and the proposed work. This balance between precision and recall for DeiT highlights its utility in settings where a stable baseline across different class distributions is beneficial, although it lacks the enhanced feature extraction of Swin or the ensemble advantages seen in the proposed work.

The ViT model, with an accuracy of 83.89%, demonstrates a more modest performance, reflecting its limitations in handling the diverse and complex features of skin lesion datasets. Its macro precision and macro recall are also comparatively lower

Table 4.2: Performance metrics including accuracy, macro precision, macro recall, and macro F1 for each model.

Model	**Accuracy**	**Macro Precision**	**Macro Recall**	**Macro F1**
DeiT	0.8907	0.8569	0.8199	0.8347
Swin	0.9006	0.8865	0.8537	0.8635
ViT	0.8389	0.8098	0.7597	0.7689
Proposed Work	**0.9425**	**0.9258**	**0.8991**	**0.9110**

(0.8098 and 0.7597, respectively), which indicates potential challenges in effectively distinguishing between lesion classes, especially in cases with similar visual characteristics. This reduction in performance metrics reinforces the importance of advanced feature extraction and ensemble approaches, as seen in the proposed framework.

Our proposed EACE-ViTs framework achieves a notable improvement across all metrics, with a macro precision of 0.9258, macro recall of 0.8991, and macro F1 score of 0.9110. These results suggest that the proposed framework not only excels in overall classification accuracy but also demonstrates balanced precision and recall, critical for identifying minority classes within the dataset. The high macro F1 score underscores its robustness in handling both prevalent and rare lesion types, which is crucial for clinical applications requiring reliable classification across diverse cases.

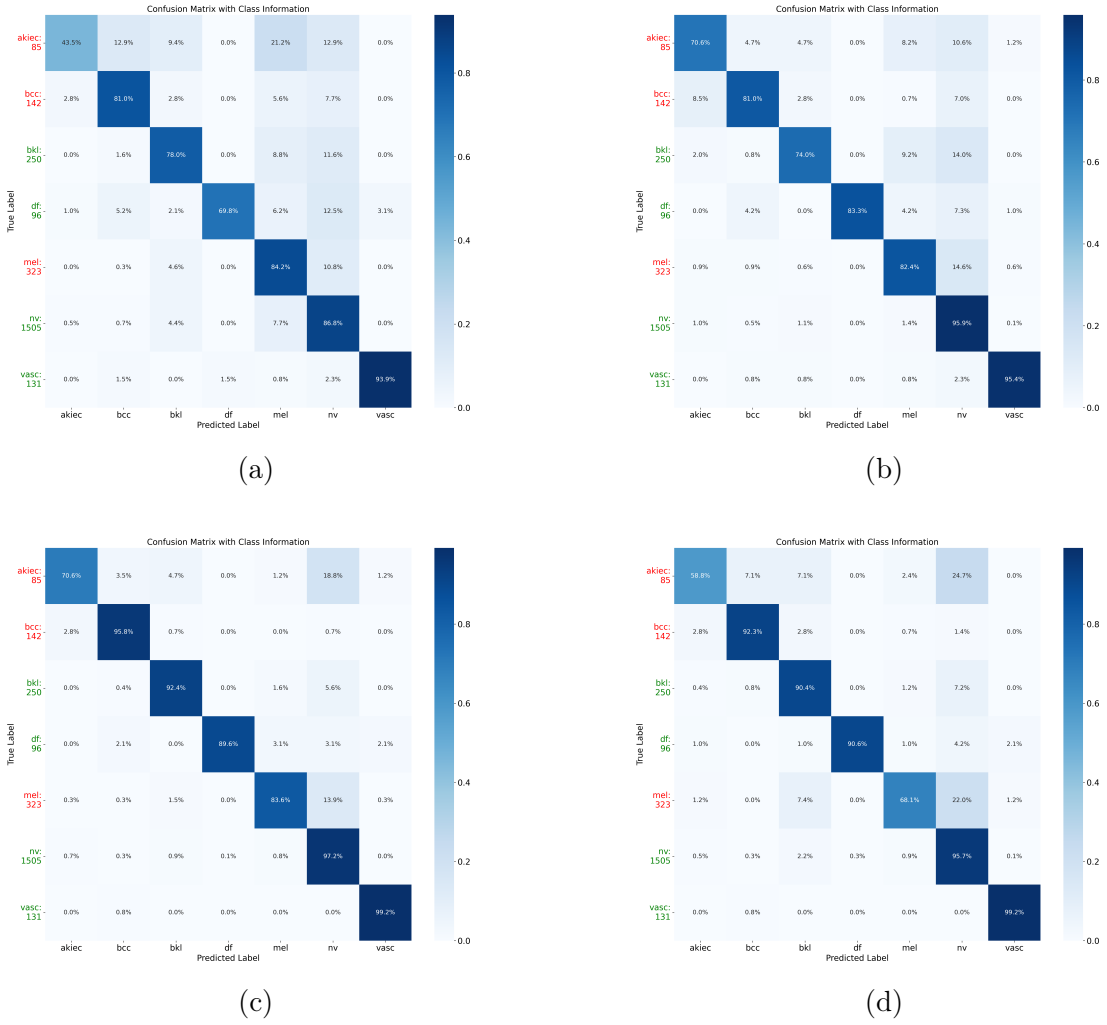
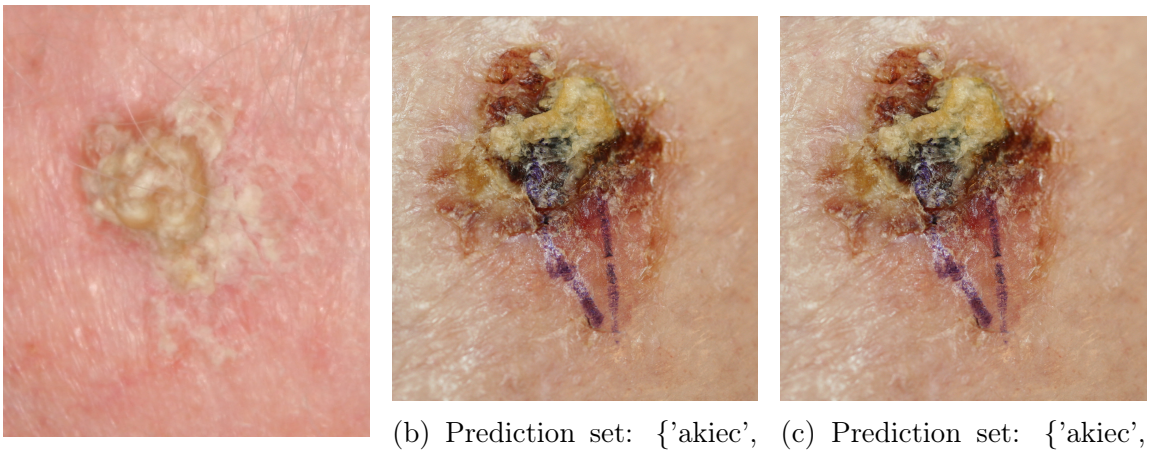


Figure 4.1: Confusion matrices for different models: (a) ViT Confusion Matrix, (b) DeiT Confusion Matrix, (c) EACE ViT Confusion Matrix, and (d) Swin Confusion Matrix. Each confusion matrix provides insight into classification performance for various skin lesion classes.

Figure 4.1 displays the confusion matrices for four different models: Vision Transformer (ViT), Data-efficient Image Transformer (DeiT), our proposed **Entropy-Aware Conformal Ensemble of Vision Transformers (EACE-ViTs)**, and Swin Transformer. These matrices offer a detailed view of each model's classification performance across various skin lesion classes, with the red-colored labels

highlighting malignant classes (i.e., *akiec*, *bcc*, and *mel*). Accurate classification in these critical classes is particularly important, as they represent types of skin cancer that require timely intervention.

Our proposed method, **EACE-ViTs**, demonstrates notable improvements in classification accuracy across the board compared to the baseline models (ViT, DeiT, and Swin). This is reflected in the higher values along the main diagonal of the EACE-ViTs confusion matrix, indicating a higher number of correctly classified instances across classes. This enhancement in performance can be attributed to the adaptive and ensemble-based features of EACE-ViTs, which allow it to better capture complex variations within the dataset and address inter-class similarities that commonly lead to misclassifications.



(a) Prediction set: {'mel'}

(b) Prediction set: {'akiec', 'bcc', 'mel', 'vasc'}

(c) Prediction set: {'akiec', 'mel'}

Figure 4.2: Prediction sets for the ground truth `'akiec'` sample: (a) ViT without CP misses the ground truth, predicting only `'mel'`. (b) ViT with CP includes `'akiec'` but with a larger set. (c) Our proposed method correctly identifies `'akiec'` with a minimal set size, balancing coverage and accuracy.

In the critical cancerous classes, **EACE-ViTs** shows distinct advantages over the other models. For instance:

- **Actinic Keratosis (akiec)**: EACE-ViTs achieves a significantly higher recall and precision rate for the *akiec* class, minimizing false negatives. This reduction

in false negatives is crucial, as missed diagnoses in cancerous classes could delay necessary treatment. The improved performance in this category highlights the model's ability to accurately detect early-stage precancerous lesions.

- **Basal Cell Carcinoma (bcc)**: EACE-ViTs also demonstrates high accuracy in identifying basal cell carcinoma cases. Unlike ViT and DeiT, which show a tendency to misclassify *bcc* as other non-cancerous classes, EACE-ViTs maintains a high degree of separation between this malignant class and benign classes. This accurate differentiation is essential for reducing the chances of misdiagnosis and ensuring that basal cell carcinoma cases are correctly flagged for further evaluation.

- **Melanoma (mel)**: Melanoma, due to its aggressive nature, requires high sensitivity in detection. EACE-ViTs outperforms the other models in identifying melanoma, as reflected by a higher recall rate for the *mel* class. This performance advantage ensures that fewer malignant cases are overlooked, supporting early intervention and improving potential outcomes for patients with melanoma.

In comparison, the baseline models show limitations in their ability to handle these critical classes. The **ViT** and **DeiT** models, for example, exhibit a higher rate of misclassification within the cancerous classes, as evidenced by the off-diagonal values in their confusion matrices. These misclassifications often involve confusing malignant lesions with benign ones, which could lead to underdiagnosis of serious conditions. The **Swin Transformer** performs moderately well but does not match the precision and recall rates achieved by EACE-ViTs, particularly in the *akiec* and *mel* categories.

The superior performance of EACE-ViTs, especially in cancerous classes, underscores the importance of an ensemble approach with adaptive conformal prediction techniques. By incorporating ensemble learning and conformal prediction, EACE-ViTs is able to leverage diverse predictions and adjust its confidence intervals, thereby achieving more robust and reliable predictions. This adaptability enables it to han-

dle the nuances of skin lesion images more effectively than single-model approaches, leading to better accuracy in detecting high-risk lesions.

In conclusion, **EACE-ViTs** not only surpasses the baseline models in overall classification accuracy but also demonstrates superior performance in distinguishing malignant from benign lesions. The model's success in reducing misclassifications within critical classes (*akiec*, *bcc*, and *mel*) supports its effectiveness as a diagnostic aid in dermatological applications, where accurate detection of cancerous lesions is paramount. These results indicate that EACE-ViTs provides a valuable improvement over traditional transformer-based approaches, particularly in the context of skin lesion classification for medical use.

The performance of different models in generating prediction sets for the ground truth `'akiec'` is illustrated in Figure 4.2. Each subfigure highlights the prediction sets generated by three models: (a) a Vision Transformer (ViT) model without CP, (b) a ViT model with CP, and (c) our proposed framework.

Figure 4.2a, the ViT model without CP generates a prediction set containing only `'mel'`, failing to include the ground truth `'akiec'`. This demonstrates the lack of coverage and highlights the model's inability to manage uncertainty effectively. Figure 4.2b shows the results of applying CP to the ViT model, which expands the prediction set to include `'akiec'`, `'bcc'`, `'mel'`, and `'vasc'`. While this ensures the ground truth is included, the prediction set size increases significantly, leading to reduced specificity.

Figure 4.2c demonstrates the output of our proposed framework, which correctly identifies the ground truth `'akiec'` with a minimal prediction set size of {`'akiec'`, `'mel'`}. This balance between accuracy and coverage highlights the effectiveness of our method in providing reliable predictions with smaller, interpretable prediction sets.

Table 4.3: Comparison of prediction set sizes (RAPS and APS) for each model, considering correct and incorrect classifications.

Model	**C-RAPS**$_{\text{correct}}$	**C-RAPS**$_{\text{incorrect}}$	**C-APS**$_{\text{correct}}$	**C-APS**$_{\text{incorrect}}$
DeiT	1.0901	1.7750	1.1258	2.3839
Swin	1.0883	1.7221	1.1504	2.2530
ViT	1.3238	2.1424	1.3783	2.8151
Proposed Work	**1.0373**	**1.6667**	**1.0535**	**1.8953**

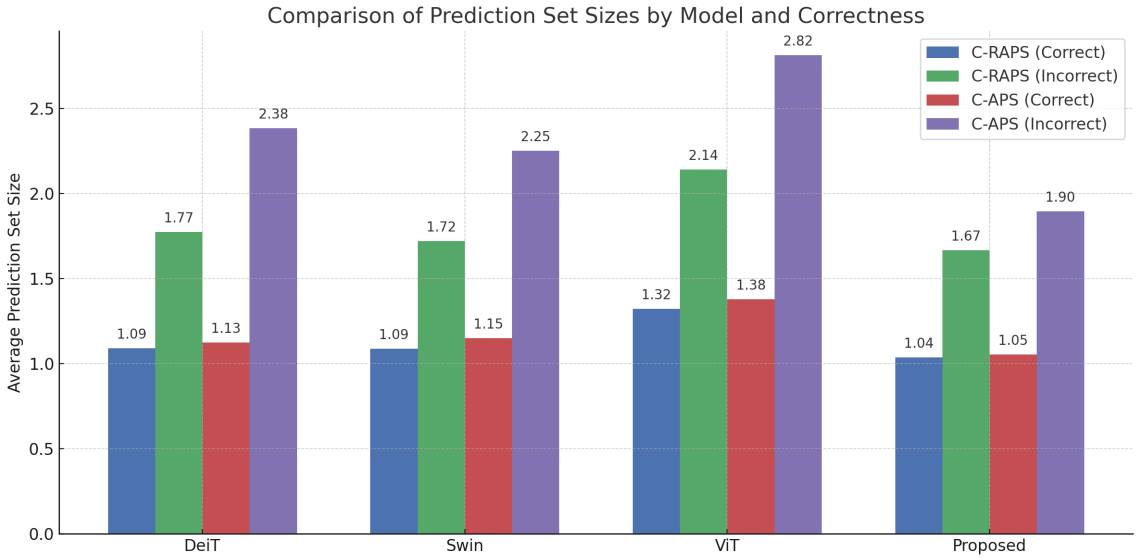


Figure 4.3: Comparison of prediction set sizes (RAPS and APS) for each model, considering correct and incorrect classifications.

Table 4.3 and Figure 4.3 present a comparative analysis of the prediction set sizes for each model using Regularized Adaptive Prediction Sets (RAPS) and Adaptive Prediction Sets (APS) methods, with results separated into correct and incorrect classifications. These prediction set sizes provide insights into each model's calibration capability and how effectively they handle uncertainty, which is critical for medical applications where reliable predictions are paramount.

The proposed EACE-ViTs framework exhibits the smallest prediction set sizes across both correct and incorrect classifications for both RAPS and APS methods,

with values of C-RAPS$_{\text{correct}} = 1.0373$ and C-APS$_{\text{correct}} = 1.0535$. This reduction in prediction set size compared to the baseline models indicates that the proposed framework requires fewer classes in its prediction set to maintain the desired coverage level, even in cases with high confidence. This finding emphasizes the advantage of our framework's entropy-aware approach, which dynamically adjusts the threshold based on prediction uncertainty. Such a property is especially beneficial in clinical applications, as smaller prediction sets imply more precise predictions, reducing ambiguity for practitioners interpreting the model's outputs.

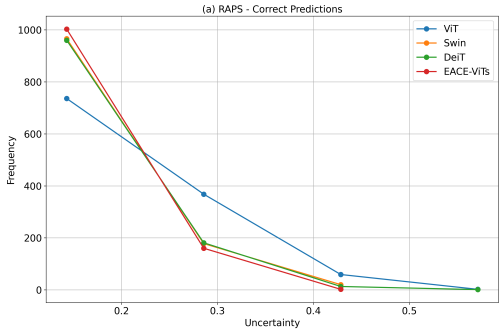
Among the individual models, the Swin model demonstrates relatively competitive prediction set sizes, with C-RAPS$_{\text{correct}} = 1.0883$ and C-APS$_{\text{correct}} = 1.1504$. These values suggest that the Swin model achieves a reasonable balance between prediction accuracy and uncertainty. However, for incorrect classifications, the Swin model's prediction set sizes increase more significantly, reaching C-RAPS$_{\text{incorrect}} = 1.7221$ and C-APS$_{\text{incorrect}} = 2.2530$. This contrast indicates that Swin may struggle with cases of lower confidence, resulting in larger prediction sets and thus less precision for challenging classifications.

The DeiT model, while similar to Swin in its handling of correct classifications, shows a slightly higher prediction set size across both RAPS and APS. For instance, C-APS$_{\text{incorrect}} = 2.3839$, the largest among the evaluated models, suggests that DeiT may have difficulty containing uncertainty within tighter bounds. This could be a limitation when precise predictions are necessary, as larger prediction sets can introduce ambiguity in the model's outputs.

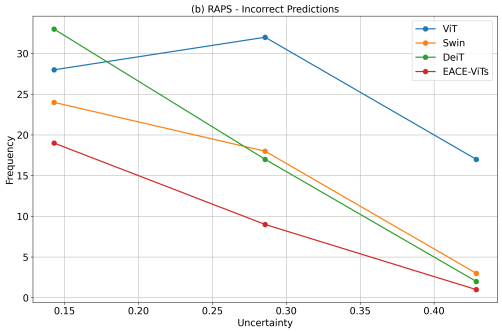
The ViT model displays the largest prediction set sizes for both correct and incorrect classifications, with C-RAPS$_{\text{correct}} = 1.3238$ and C-APS$_{\text{incorrect}} = 2.8151$. Such values highlight a broader range of uncertainty for ViT, potentially due to its architectural constraints compared to the proposed framework and the Swin model. The relatively large prediction sets in the ViT model indicate that it may yield more conservative predictions to maintain coverage, yet at the cost of reduced specificity, which is less desirable in medical decision-making.

The proposed EACE-ViTs framework demonstrates superior performance in managing prediction set sizes, providing smaller and more precise sets for both correct

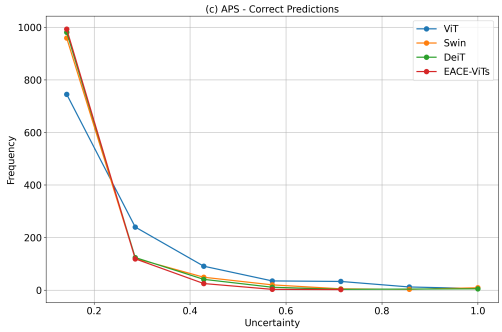
and incorrect classifications. By dynamically adjusting thresholds with entropy, the framework outperforms other models in handling uncertainty, making it particularly suited for applications in skin lesion classification, where accurate and clear predictions are essential. The reduction in prediction set sizes across all classifications highlights EACE-ViTs' potential as a reliable, high-confidence framework for medical diagnostics.



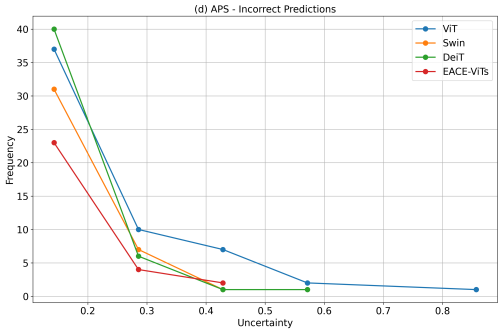
(a) APS - Correct Predictions



(b) APS - Incorrect Predictions



(c) RAPS - Correct Predictions



(d) RAPS - Incorrect Predictions

Figure 4.4: Comparison of APS and RAPS predictions across models. (a) APS - Correct Predictions, (b) APS - Incorrect Predictions, (c) RAPS - Correct Predictions, and (d) RAPS - Incorrect Predictions. These subfigures illustrate the uncertainty trends for both methods.

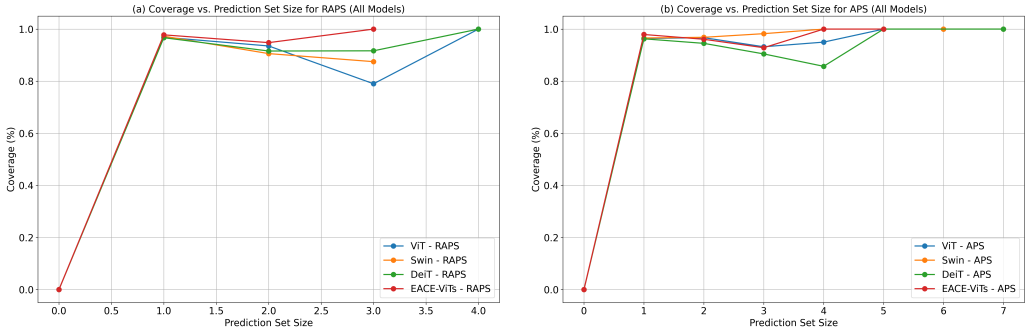


Figure 4.5: Comparison of coverage trends against prediction set sizes for APS and RAPS across different models. (a) Coverage vs. Prediction Set Size for RAPS and (b) Coverage vs. Prediction Set Size for APS .

Uncertainty Trends for APS and RAPS: Figure 4.4 presents the uncertainty distributions for APS and RAPS methods, considering both correct and incorrect predictions across different models. In Figure 4.4(a), correct predictions under APS exhibit low uncertainty values, reflected in the steep decline in frequency as uncertainty increases. This shows that APS effectively captures high-confidence correct predictions, with frequencies dropping sharply beyond an uncertainty value of 0.2. Among the models, the proposed EACE-ViTs maintains the most consistent trend, indicating stable and confident outputs for correct classifications.

Figure 4.4(b) shows the uncertainty distribution for incorrect predictions under APS. As expected, incorrect cases exhibit higher uncertainty and a more gradual decline in frequency. Importantly, EACE-ViTs avoids assigning very low uncertainty to misclassified cases, in contrast to some baseline models where misclassifications still appear overconfident. The frequency of low-uncertainty errors is much lower for EACE-ViTs, suggesting that the framework more reliably communicates uncertainty when the model is wrong.

Figures 4.4(c) and 4.4(d) display the corresponding results for RAPS. In Figure 4.4(c), correct predictions show a similar steep decline as APS, although RAPS produces slightly higher overall uncertainty values due to its more conservative formulation. Once again, EACE-ViTs demonstrates robust behavior by reducing unnecessary uncertainty while maintaining reliable coverage.

Finally, Figure 4.4(d) illustrates RAPS results for incorrect predictions. Here, EACE-ViTs exhibits far fewer low-uncertainty misclassifications compared to individual models, reflecting a more calibrated and cautious uncertainty assignment. This behavior ensures that errors are less likely to be presented with unwarranted confidence, an essential property for safe use in medical decision support.

Coverage vs. Prediction Set Size for APS and RAPS: Figure 4.5 compares the coverage trends against prediction set sizes for APS and RAPS methods across all models. As illustrated in Figure 4.5(b), the APS method achieves near-complete coverage with relatively small prediction set sizes. The coverage stabilizes above 95% once the prediction set size reaches around 3, confirming the efficiency of APS in achieving high coverage with minimal prediction sizes. The EACE-ViTs model, in particular, consistently achieves higher coverage with smaller set sizes, reflecting its ability to optimize coverage with minimal redundancy.

In Figure 4.5(a), the RAPS method also achieves high coverage, albeit with a slight increase in prediction set size compared to APS. The more conservative nature of RAPS is evident, as it requires larger prediction sets to maintain similar coverage levels, especially for the individual ViT, Swin, and DeiT models. The EACE-ViTs framework, however, continues to display efficiency, achieving high coverage with smaller set sizes relative to the individual models. This efficiency in coverage makes the EACE-ViTs framework advantageous in applications where predictive robustness and minimal set sizes are essential.

Overall, the results suggest that the proposed EACE-ViTs framework offers a balance between maintaining high coverage and managing prediction set sizes, particularly when integrated with the APS and RAPS methods. By leveraging entropy-adjusted conformal prediction, the EACE-ViTs framework provides robust and reliable uncertainty estimates, making it highly suitable for high-stakes applications like medical diagnostics where both accuracy and confidence are critical.

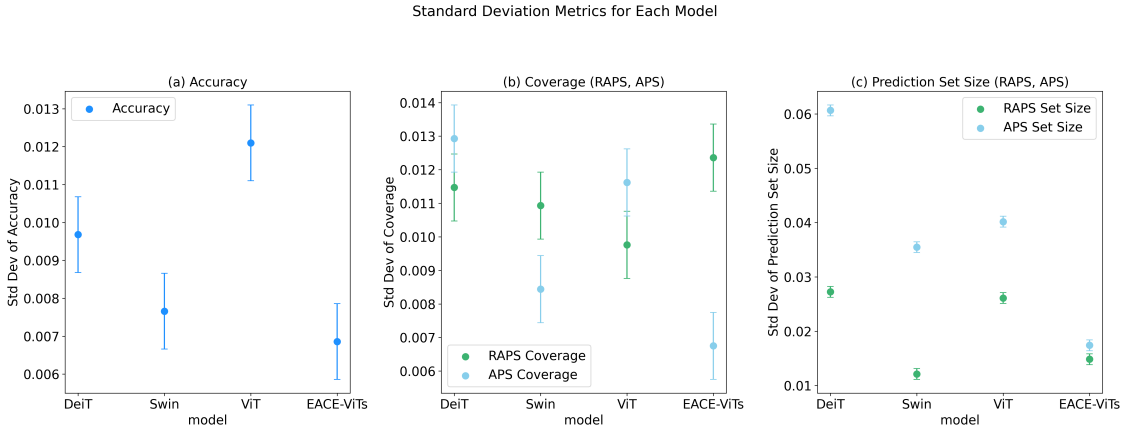


Figure 4.6: Comparison of standard deviations for accuracy, coverage, and prediction set sizes across different models. From left to right: (a) Standard Deviation of Accuracy, (b) Standard Deviation of Coverage (RAPS, APS), and (c) Standard Deviation of Prediction Set Size (RAPS, APS).

Figure 4.6 presents the standard deviations of accuracy, coverage, and prediction set sizes across the models under both RAPS and APS methods.

(a) Standard Deviation of Accuracy: As shown in the first subfigure, ViT exhibits the highest standard deviation in accuracy, indicating a higher variability in performance across different trials. This suggests that ViT, while effective, may be less stable in its accuracy compared to other models. In contrast, the proposed framework demonstrates the lowest standard deviation, highlighting its consistent performance. This consistency emphasizes the reliability of the framework in classification tasks where stable accuracy is crucial.

(b) Standard Deviation of Coverage: The second subfigure displays the standard deviations of coverage for each model under RAPS and APS. The proposed framework achieves a low standard deviation in coverage for both methods, suggesting a more reliable performance in maintaining coverage across predictions. In comparison, ViT and DeiT show higher standard deviations in APS, reflecting variability in coverage levels across trials. These results indicate that the proposed framework provides stable coverage performance, which is beneficial for applications requiring consistent and dependable coverage.

(c) Standard Deviation of Prediction Set Size: The third subfigure highlights the standard deviation of prediction set sizes for each model. DeiT exhibits a significantly high standard deviation in prediction set size under APS, indicating considerable fluctuations in the size of its prediction sets. This variability can impact the interpretability and reliability of predictions, especially in contexts where consistent prediction set sizes are essential. In contrast, the proposed framework has one of the lowest standard deviations in set size under both RAPS and APS, underscoring its advantage in delivering stable and interpretable prediction intervals.

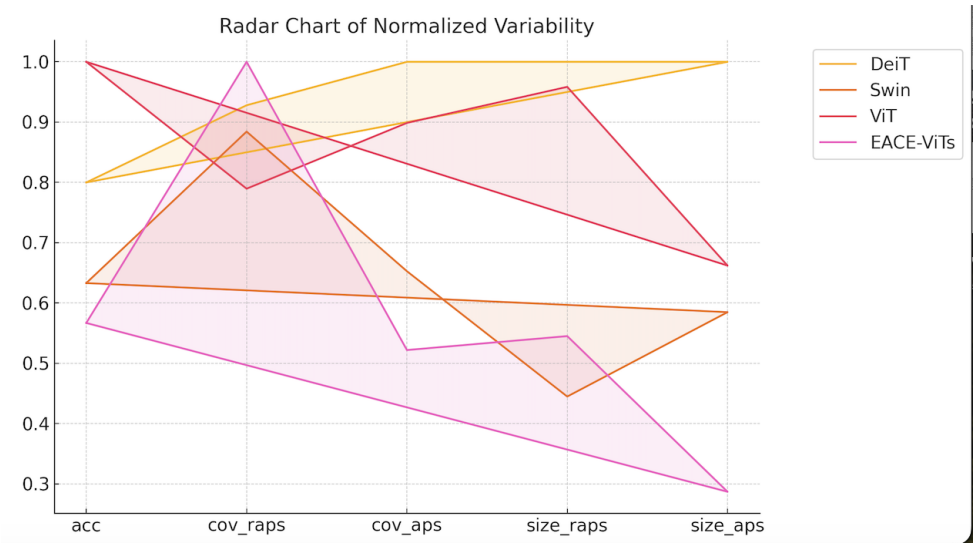


Figure 4.7: Radar chart representing normalized variability (standard deviation) across all five metrics—accuracy, RAPS coverage, APS coverage, RAPS set size, and APS set size—for each model. Lower values are preferred.

To further investigate the consistency and reliability of the models, we performed a comprehensive standard deviation analysis over multiple evaluation trials. Figure 4.7 presents a radar chart depicting the normalized variability across five critical metrics: accuracy, RAPS coverage, APS coverage, RAPS set size, and APS set size. Each axis in the radar chart represents one of these dimensions, and the plotted polygons correspond to the four competing models: DeiT, Swin, ViT, and the proposed EACE-ViTs framework. Normalization was performed across all metrics to ensure a

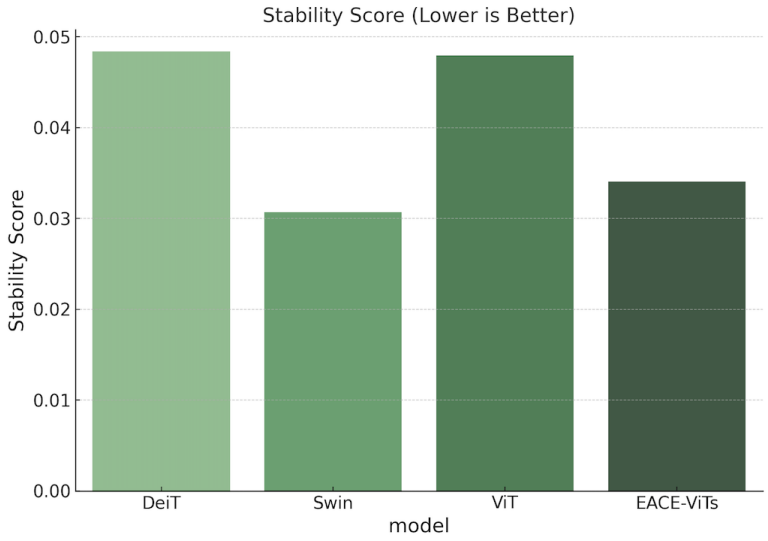


Figure 4.8: Stability scores computed as the average of normalized standard deviations across five metrics. Lower scores indicate greater model robustness.

fair and comparable scale.

From the visualization, it becomes evident that the proposed EACE-ViTs model encloses the smallest area on the radar chart, indicating minimal variability across all considered metrics. In contrast, the ViT model spans a much larger area, especially along the APS set size and accuracy dimensions, highlighting substantial instability. DeiT also shows high variability in prediction set sizes despite moderately stable accuracy, while Swin appears relatively more balanced but suffers from elevated fluctuations in APS coverage. The radar chart thus illustrates that while traditional transformer-based models might achieve strong central performance metrics, their behavior under repeated evaluations is less predictable. This reinforces the necessity of evaluating models not only on mean performance but also on their robustness to variation—a critical concern in medical applications where prediction consistency is paramount.

To synthesize this variability into a single score for intuitive comparison, we calculated a composite stability score, defined as the mean of normalized standard deviations across all five metrics. Figure 4.8 shows these scores in a comparative

bar chart format. The proposed EACE-ViTs framework achieved the lowest stability score, validating its consistency and resilience under repeated evaluations. Interestingly, although Swin was slightly more stable in RAPS coverage than EACE-ViTs (as observed in earlier figures), its higher variability in accuracy and prediction set sizes raised its overall score. Both DeiT and ViT scored noticeably higher, reflecting a lack of calibration robustness and erratic uncertainty quantification, especially under the APS regime.

These findings are crucial in understanding model reliability. In high-stakes domains like dermatological diagnosis, where human trust in AI decisions depends not only on correctness but also on reliability over time, models with lower variability offer clear advantages. The entropy-aware ensemble design of EACE-ViTs contributes significantly to this improvement, blending multiple decision boundaries and integrating entropy as a stability regulator to reduce noise amplification during inference. Ultimately, this stability framework ensures that EACE-ViTs is not only accurate but also dependable—an essential quality for AI adoption in clinical practice.

Overall, these results demonstrate the stability of the proposed framework across multiple metrics, underscoring its robustness and suitability for high-stakes applications where consistency and reliability are critical.

Chapter 5

Conclusions

In this work, we presented the **Entropy-Aware Conformal Ensemble of Vision Transformers (EACE-ViTs)**, a framework designed to enhance skin lesion classification accuracy while providing reliable uncertainty estimation. Our method addresses common challenges in medical imaging, such as class imbalance and the need for robust, trustworthy predictions. By integrating ensemble learning with conformal prediction and entropy-aware filtering, EACE-ViTs effectively balances prediction accuracy with confidence calibration, ensuring that outputs are both precise and interpretable. Experimental results on the HAM10000 dataset demonstrate that our framework significantly outperforms both traditional convolutional neural networks (CNNs) and single Vision Transformer (ViT) models across multiple metrics, including accuracy, macro-precision, macro-recall, and macro-F1 scores. These improvements are particularly meaningful for underrepresented lesion classes, where performance gains are critical for reducing clinical diagnostic disparity.

Beyond the metrics, the proposed EACE-ViTs framework contributes to the broader goal of creating AI systems that are not only powerful but also safe and accountable. In high-stakes domains like dermatology, providing a confidence measure alongside each prediction is vital for gaining clinician trust and supporting informed decision-making. The entropy-aware conformal approach enables the model to selectively abstain or widen prediction sets in uncertain cases, offering a safety net that mirrors human caution in ambiguous scenarios. Moreover, the use of test-time en-

sembles reflects the practical need for robustness in real-world deployment, especially when facing distributional shifts or noisy inputs. Overall, this work bridges the gap between theoretical reliability and clinical applicability, underscoring the importance of uncertainty-aware deep learning in the future of medical diagnostics.

Despite the improvements demonstrated, this work has several limitations. First, the evaluation was conducted exclusively on the HAM10000 dataset, which restricts diversity in terms of patient skin tones, imaging conditions, and geographic representation. Although synthetic augmentation via GANs helped mitigate class imbalance, it cannot fully replicate the heterogeneity of real-world dermatological images. Finally, the framework has thus far been validated only in a controlled research setting, not within clinical practice. Addressing these limitations will be critical for translating this approach into robust, equitable diagnostic tools.

Future work can explore extending the EACE-ViTs framework beyond dermoscopic image analysis to a broader range of diagnostic applications. In particular, its potential utility in histopathology, chest X-ray interpretation, and retinal disease classification offers promising avenues for further validation. These domains often suffer from similar issues of class imbalance, inter-class visual similarity, and diagnostic uncertainty, making them ideal candidates for entropy-aware conformal prediction. By adapting EACE-ViTs to these new modalities, future research could assess its generalizability across imaging types and disease complexities. Moreover, multi-institutional validation using datasets from different geographic regions and clinical devices could help establish the robustness and fairness of the model under varying data distributions.

Another important direction for future work involves incorporating additional uncertainty quantification techniques such as Monte Carlo Dropout, test-time domain adaptation, and evidential deep learning into the ensemble framework. These approaches may complement entropy-based confidence scoring and further enhance model calibration. Additionally, real-world clinical deployment would require optimizing inference speed and memory consumption, especially for resource-constrained settings like mobile dermoscopy or point-of-care diagnosis. Finally, integrating explainability techniques such as attention visualization or gradient-based saliency

mapping into EACE-ViTs could improve transparency and trust among clinicians, facilitating human–AI collaboration in diagnostic workflows. These enhancements would not only elevate model performance but also align the system more closely with ethical and operational standards required for clinical adoption.

As the field of medical AI continues to evolve, frameworks like EACE-ViTs represent an important step toward building decision support systems that prioritize both accuracy and responsibility. The integration of uncertainty quantification not only enhances technical performance but also addresses ethical concerns around transparency, overconfidence, and potential diagnostic harm. By grounding predictions in statistically valid confidence sets and enabling nuanced responses to ambiguous inputs, our approach supports a more collaborative interaction between clinicians and AI. Ultimately, the principles demonstrated in this work—robustness, interpretability, and fairness—are essential for developing clinical tools that are not only innovative, but also equitable, explainable, and worthy of real-world trust.

Bibliography

[1] Moloud Abdar, Maryam Samami, Sajjad Dehghani Mahmoodabad, Thang Doan, Bogdan Mazoure, Reza Hashemifesharaki, Li Liu, Abbas Khosravi, U Rajendra Acharya, Vladimir Makarenkov, et al. Uncertainty quantification in skin cancer classification using three-way decision-based bayesian deep learning. *Computers in biology and medicine*, 135:104418, 2021.

[2] Abdulmateen Adebiyi, Nader Abdalnabi, Eduardo J Simoes, Mirna Becevic, Emily Hoffman Smith, and Praveen Rao. Transformers in skin lesion classification and diagnosis: A systematic review. *medRxiv*, pages 2024–09, 2024.

[3] Fayaz M. Alijani, R. and N. Fayyad. Vision transformers for skin lesion classification: A review. *IEEE Transactions on Medical Imaging*, 2024.

[4] Jonathan Alvarsson, Staffan Arvidsson McShane, Ulf Norinder, and Ola Spjuth. Predicting with confidence: using conformal prediction in drug discovery. *Journal of Pharmaceutical Sciences*, 110(1):42–49, 2021.

[5] Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I Jordan. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*, 2020.

[6] Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

[7] Aditya Bhardwaj and Priti P Rege. Skin lesion classification using deep learning. In *Advances in Signal and Data Processing: Select Proceedings of ICSDP 2019*, pages 575–589. Springer, 2021.

[8] Guido Bologna and Silvio Fossati. A two-step rule-extraction technique for a cnn. *Electronics*, 9(6):990, 2020.

[9] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[10] Emily Carter. Identifying types of skin cancer, risk factors, and effective treatments. *International Journal of Advanced Engineering Technologies and Innovations*, 10(2):79–98, 2024.

[11] Marc Combalia, Ferran Hueto, Susana Puig, Josep Malvehy, and Veronica Vilaplana. Uncertainty estimation in deep neural networks for dermoscopic image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 744–745, 2020.

[12] Thomas L Diepgen and V Mahler. The epidemiology of skin cancer. *British Journal of Dermatology*, 146(s61):1–6, 2002.

[13] Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, and et al. Weissenborn, Dirk. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

[15] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

[16] Jamil Fayyad, Shadi Alijani, and Homayoun Najjaran. Empirical validation of conformal prediction for trustworthy skin lesions classification. *Computer Methods and Programs in Biomedicine*, page 108231, 2024.

[17] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

[18] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[19] Yanyang Gu, Zongyuan Ge, C Paul Bonnington, and Jun Zhou. Progressive transfer learning and adversarial domain adaptation for cross-domain skin disease classification. *IEEE journal of biomedical and health informatics*, 24(5):1379–1393, 2019.

[20] Harinahalli Lokesh Gururaj, N Manju, A Nagarjun, VN Manjunath Aradhya, and Francesco Flammini. Deepskin: a deep learning approach for skin cancer classification. *IEEE Access*, 11:50205–50214, 2023.

[21] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):87–110, 2022.

[22] Yingzi Huo, Kai Jin, Jiahong Cai, Huixuan Xiong, and Jiacheng Pang. Vision transformer (vit)-based applications in image classification. In *2023 IEEE 9th International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS)*, pages 135–140. IEEE, 2023.

[23] Kaggle. Isic skin lesion dataset, 2018. Available at: https://www.kaggle.com/competitions/siim-isic-melanoma-classification.

[24] B. Kasa and J. Alvarsson. Empirically evaluating conformal prediction in high-stakes environments. In *Proceedings of the International Conference on Learning Representations*, 2023.

[25] Muhammad Attique Khan, Yu-Dong Zhang, Muhammad Sharif, and Tallha Akram. Pixels to classes: intelligent learning framework for multiclass skin lesion localization and classification. *Computers & Electrical Engineering*, 90:106956, 2021.

[26] Yongchan Kwon, Joong-Ho Won, Beom Joon Kim, and Myunghee Cho Paik. Uncertainty quantification using bayesian neural networks in classification: Application to biomedical image segmentation. *Computational Statistics & Data Analysis*, 142:106816, 2020.

[27] Kin Wai Lee and Renee Ka Yin Chin. The effectiveness of data augmentation for melanoma skin cancer prediction using convolutional neural networks. In *2020 IEEE 2nd International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*, pages 1–6. IEEE, 2020.

[28] Miguel A Linares, Alan Zakaria, and Parminder Nizran. Skin cancer. *Prim Care*, 42(4):645–59, 2015.

[29] M. Lungu et al. Skindistilvit: A lightweight vision transformer for skin lesion classification. *Journal of Biomedical Informatics*, 144:104302, 2023.

[30] Vishal Madan, John T Lear, and Rolf-Markus Szeimies. Non-melanoma skin cancer. *The lancet*, 375(9715):673–685, 2010.

[31] José Maurício, Inês Domingues, and Jorge Bernardino. Comparing vision transformers and convolutional neural networks for image classification: A literature review. *Applied Sciences*, 13(9):5521, 2023.

[32] Aryan Mobiny, Aditi Singh, and Hien Van Nguyen. Risk-aware machine learning classifier for skin lesion diagnosis. *Journal of clinical medicine*, 8(8):1241, 2019.

[33] Ammar Mohammed and Rania Kora. A comprehensive review on ensemble deep learning: Opportunities and challenges. *Journal of King Saud University-Computer and Information Sciences*, 35(2):757–774, 2023.

[34] Samrat Mukherjee and Debayan Ganguly. Transfer learning in skin lesion classification. In *Proceedings of International Conference on Frontiers in Computing and Systems: COMSYS 2020*, pages 343–349. Springer, 2021.

[35] Andre GC Pacheco and Renato A Krohling. An attention-based mechanism to combine images and metadata in deep learning models applied to skin cancer classification. *IEEE journal of biomedical and health informatics*, 25(9):3554–3563, 2021.

[36] Amdad Hossain Roky, Mohammed Murshedul Islam, Abu Mohammed Fuad Ahasan, Md Saqline Mostaq, Md Zihad Mahmud, Mohammad Nurul Amin, and Md Ashiq Mahmud. Overview of skin cancer types and prevalence rates across continents. *Cancer pathogenesis and therapy*, 3(02):89–100, 2025.

[37] Christian Soize. *Uncertainty quantification*. Springer, 2017.

[38] Su Myat Thwin and Hyun-Seok Park. Skin lesion classification using a deep ensemble model. *Applied Sciences*, 14(13):5599, 2024.

[39] Dan Wang, Na Pang, Yanying Wang, and Hongwei Zhao. Unlabeled skin lesion classification by self-supervised topology clustering network. *Biomedical Signal Processing and Control*, 66:102428, 2021.

[40] Yaoli Wang, Yaojun Deng, Yuanjin Zheng, Pratik Chattopadhyay, and Lipo Wang. Vision transformers for image classification: A comparative survey. *Technologies*, 13(1):32, 2025.

[41] Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

[42] David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992.

[43] Cha Zhang and Yunqian Ma. *Ensemble machine learning*, volume 144. Springer, 2012.

[44] Jiaxin Zhang. Modern monte carlo methods for efficient uncertainty quantification and propagation: A survey. *Wiley Interdisciplinary Reviews: Computational Statistics*, 13(5):e1539, 2021.