National College of Ireland

Risk Adaptive Triple Stream Vision Transformer for Dermatological Triage

MSc Research Project
Master of Science in Data Analytics

Saran Das Sivadasan Achary
Student ID: 23336030

School of Computing
National College of Ireland

Supervisor: Athanasios Staikopoulos

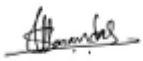
National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Saran Das Sivadasan Achary
Student ID:	23336030
Programme:	Master of Science in Data Analytics **Year:** 2025
Module:	MSc Practicum Part 2
Supervisor:	Athanasios Staikopoulos
Submission Due Date:	28/01/2026
Project Title:	Risk Adaptive Triple Stream Vision Transformer for Dermatological Triage
Word Count:	8075 **Page Count 21**

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: 

Date: 26/01/2026

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Risk Adaptive Triple Stream Vision Transformer for Dermatological Triage

Saran Das Sivadasan Achary
23336030

Abstract

Skin diseases are a health care challenge especially in areas where dermatologists are not readily accessible. As much as computer-assisted triage systems hold promise in ranking high-risk cases, the existing approaches currently have pre-trained model bias and are missing such critically important clinical characteristics as calibration and interpretability. Our suggested model is a Risk-Adaptive Triple-Stream ViTra (TS-ViT) which trains on a dataset of dermatological images to identify lesions into the categories of HIGH, MODERATE, and LOW clinical urgency. Its architecture combines three streams, namely global classification to obtain whole-view risk assessment, spatial attention to obtain interpretable heatmaps and uncertainty quantification with Monte Carlo dropout to estimate confidence. TS-ViT (12,345 dermatoscopic images, 40 subclasses) demonstrated accuracy (89.3%) and macro F1-score (0.847), which is significantly less than ViT-B/16 (67.2%), ResNet-50 (63.8%), and MobileNetV3-Large (59.4%). The model had better calibration (ECE: 0.074, Brier Score: 0.098), and good uncertainty-error correlation (Spearman $r = 0.68$, $p < 0.001$), which allowed the use of triage policies based on confidence. There was correspondence between the spatial attention maps and the diagnostic criteria, with irregular borders and asymmetric pigmentation being significantly observed in the malignant lesions. Such findings indicate that multi-stream architectures that are cautious of risks are superior to classical accuracy-based models in clinical utility, which is the base of safe and interpretable dermatological triage in constraint resources.

Keywords: Triple-Stream Vision Transformer, Skin Lesion Classification, Interpretable Deep Learning, Dermatological Triage, Uncertainty Quantification, Model Calibration

1 Introduction

The skin cancers and related dermatological diseases contribute above 30 percent of the entire cancers cases that have been diagnosed throughout the world. The curative process of skin cancer would be successful provided the cancer is caught early in life. Since most of the countries lack the services of dermatologists especially in the rural setting, the timely diagnosis of skin cancer has been an issue. The necessity to present machine searching methods toward automated detection of skin cancer has, therefore, emerged.

CNNs, as well as other deep learning-based applications, achieve the diagnostic accuracy of a dermatologist on properly curated datasets. However, the performance of a model in a controlled setting is not a sufficient measure in terms of diversification in a real-world clinical practice. In order to truly serve in a medical capacity, a medical AI solution must be able to

present its levels of confidence in a more palatable and implementable way to a human clinician managing the AI system. Today in the field of medical AI used in dermatology, there are three fundamentally missing items, namely: First, CNNs are unaware of global representations of visual context and they depend only on local representations. Secondly, there is the impact of ImageNet pretraining to get a biased model because of the resultant space of features representation that are not well correlated to dermatological morphology-based features representation in the first place. Thirdly, the prediction of confidence of the model is usually not well calibrated thus giving unreliable estimates of certainty in making clinical decisions. Vision Transformers resolve the problems of global context and feature representation by using the self-attention mechanism. Nevertheless, the current methods still use transfer learning of non-medical fields. In this study TS-ViT (Risk-Adaptive Triple-Stream Vision Transformer) is presented, a brand new architecture that has been completely trained out of scratch on dermatology data, which combines global classification, spatial attention and uncertainty quantification in order to achieve safe skin lesions triage in clinics.

1.1 Research Question

Can a Vision Transformer train from scratch with risk-aware attention mechanisms effectively classify skin lesions according to their clinical urgency while offering spatial explanations and calibrated confidence estimates?

1.2 Research Objectives

There are some research objectives in this study which are:

1. Train a Vision Transformer model directly on the DERM12345 dataset (12,345 dermatoscopic images) without ImageNet pretraining to eliminate non-medical feature bias. Utilize multi-task learning by combining weighted cross-entropy loss, entropy-concentration regularization, and Brier score optimization to jointly train global classification, spatial attention localization, and uncertainty estimation within a unified transformer architecture.
2. Compare performance against baseline architectures (ResNet-50, vanilla ViT-B/16, and MobileNetV3-Large) under same experimental conditions across multiple evaluation metrics: classification accuracy, per-class performance addressing class imbalance, and calibration quality measured by Expected Calibration Error and Brier score.
3. Demonstrate how uncertainty estimates facilitate effective clinical triage: high-confidence predictions proceed with minimal intervention while low-confidence predictions are flagged for specialist review, satisfying requirements for safe medical AI deployment.

This research is divided into seven parts. Section 1 presents the background, research question, objectives, and the importance of developing a risk-adaptive vision transformer to develop dermatological triage. Section 2 conducts a related work review of skin disease classification

with the deep learning approach, vision transformers in medical imaging, hybrid architecture, uncertainty measure, and multi-task learning, addressing three key research gaps. Section 3 shows the research approach with the selection of the dataset, preprocessing plans, the infrastructure of the experiment, and the architecture of the baseline. Section 4 lists the design specification of the TS-ViT with a description of the triple-stream design, the baseline models, and the multi-task learning structure. Section 5 provides information about the implementation that includes development of environment, data processing pipeline, model integration and training procedures. Section 6 measures performance by classification accuracy, calibration analysis, visualization of spatial attention, quantification of uncertainty and ablation analysis. Section 7 is a conclusion, clinical implications discussion and discussion of future work directions.

2 Related Work

2.1 Classification of Skin Disease Using Deep Learning

The past ten years have been characterized by an outstanding presence in the dermatological industry because of the use of deep learning. In their article in 2017, Esteva et al. demonstrated that under the conditions of large volumes of data, CNNs can accurately classify skin cancer at the level of human dermatologists. Success of the Esteva work has been used as a major platform by the work of Brinker et al. in 2019 whereby the findings showed that deep learning networks were equivalent to dermatologists and higher skills in detecting melanoma. However, the problems of domain shift that appear when the ImageNet is applied to such models are always present. Another possible source of bias in deep learning applications to dermatological problems is that most use pretrained models generated by the ImageNet project during which non-medical features tend to be represented. Although the progress in the use of deep learning in image classification tasks has been successful, ImageNet pretraining presents a challenge when it comes to domain-specific dermatological diagnosis.

2.2 Vision Transformers in Medical Imaging

Vision Transformers (ViT) models have addressed multiple weaknesses of convolutional networks by introducing the potential of holistic image processing by implementing the self-attention mechanism. The authors Chen et al. (2022) came up with the term TransFusion and used multi-scale feature fusion method to improve the outcomes of histopathology image classification tasks compared to CNNs. Liu et al. (2023) developed the Swin Transformer V2 model whereby the hierarchical attention windows concept was used to support processing diverse high-resolution images. The authors of the study (Li et al., 2024) made use of the LViT model to increase the multimodal learning process by adding to the results of the medical image segmentation problems text embedding. The efficacy of following the ViT model in a self-supervised fashion to deliver the complete results in the procedure of diagnosing brain tumors based on the usage of different imaging techniques was well demonstrated by Chen et al. (2024). Kundu et al. (2024) detailed that the state-of-the-art performance was reached by means of applying DINOv2-based Vision Transformers and a rather small set of MRIs are being used.

The most important disadvantage of ViT models implementation, however, consists in the fact that nearly every existing approach to the ViT models implementation relies on ImageNet or CLIP-pretrained backbones, which may introduce non-medical feature bias into the clinical settings.

2.3 Hybrid and Multi-Scale Transformer Architectures

The creation of hybrid models with CNNs and transformer designs has just recently taken place to absorb the benefits of learning local features and computing global context in way that is mutually advantageous. Kim, Khan, and Banerjee (2024) conducted a systematic review of the use of hybrid models in the radiology field and found that the accuracy of the diagnosis showed an increase along with a subsequent rise in the cost of computation. Hu et al. (2025) developed a multi-scale ViT model to predict diseases to benefit the lesion-level of information and the global context information is retained in a robust manner. Yang et al. (2025) created a ViT architecture that integrates radiomic values to bring quantitative imaging biomarkers with visual representation learning to achieve efficient medical image interpretation. The recent instances of models are the implementation of the MedMoE-ViTs model relying on the mixture-of-experts learning techniques proposed to different modalities within the medical images context. Although there has been an increment in model complexities and costs in computing, the current interest lay in enhancing prediction efficacy of models. But the recent progress in the field places more importance on predictive performance than clinical explainability, and medical practitioners are unable to have a solid reason to make critical model predictions.

2.4 Uncertainty Quantification and Interpretability

The prediction models should be accurate and well-calibrated in the estimates of confidence. Abdar et al. (2021) used a Monte Carlo dropout strategy to achieve quantification of uncertainty in deep learning the first to describe this, shortly preceded by Gal and Ghahramani (2016) to quantify uncertainty in an analysis of skin cancers, and demonstrate the situations where uncertainty relates to the probability of error in clinical treatment when applied on a sample population. These studies however highlight that most of the existing techniques of quantifying uncertainty are post-hoc techniques which do not rely on the model training process and therefore generate ill calibrated confidence estimates. ViT attention maps provide an explanatory ability related to the focus of attention in the model but provide no measurement process of uncertainty coupled to the process of clinical risk assessment. This prediction-explanation-confidence asymmetry remains one of the key constraints of the present medical AI applications.

2.5 Gap Analysis and Research Motivation

There is much development in AI in dermatology, but in the methods provided there are three significant limitations and the current practice is planned to address each of them.

Gap 1: Domain Shift due to ImageNet Pretraining

Typically, the existing dermatology models utilizing the current transformer are investigated on ImageNet-pretrained backbone models (Chen et al., 2022; Liu et al., 2023; Li et al., 2024). Recent discoveries propose that the convergence rate is shorter actually with ImageNet pretraining but encourages better counts of overfitting with validation losses at 16.67% compared to -33.33%. According to the study by Daneshjou et al. (2022), authors have discovered that there is more disparity in the degree of performance among individuals of different skin tones when they utilized ImageNet-pretrained models. It lacks any researches to confirm whether one requires datasets of over 10,000 dermatoscopic images to be able to use Vision Transformers with no extra training and get the best results without the influence of non-medical features.

Gap 2: Decoupled Interpretability and Calibration
Nonetheless, the available models are unable to provide spatial explainability and confidence calibration at the same time. Both the significance of attention maps based on clinically valuable regions and the absence of the calibration of confidence in prediction tasks are also addressed in the study conducted by Zhang et al. (2021), in which the attention maps do not contain any of the measures of calibration of the assurance of the predictions. The investigation by Abdar et al. (2021) to classify skin cancers aimed at pointing out the difficulty behind models in their approach of intuiting uncertainty calculations, thus necessitating the deliberate application of calibration methods. This disconnection renders the use of clinical applications difficult as the attention maps serve to explain the position of a model where it is looking without providing any information about the stability of that evaluation. The joint solution of spatial reasoning and confidence calibration would be the most appropriate assistance to clinicians in risk-stratified triaging tasks.

Gap 3: Post-Hoc vs. Integrated Uncertainty Quantification
Uncertainty quantification as a design factor has been not addressed as much in the design of architecture. When Gal and Ghahramani (2016) used Monte Carlo Dropout and applied it to dermatology (Abdar et al., 2021), random sampling is used in predicative machinery; however, it has no impact on the learning-of-feature space aspect of the model. The usage in multi-task learning is of those architectures that process various kinds of tasks like in the models designed by Hu et al. (2025) and Yang et al. (2025), but they do not learn it in the model training process. Systems integrating the joint optimization of classification, attention, and uncertainty quantification to make confidence-based clinical referral decisions design architectures are a gap unexplored.

The TS-ViT model fills all these gaps as follows:

(1) The model is trained on 12345 dermatoscopic images of the DERM12345 dataset with no ImageNet pretraining to show that it is able to learn domain-specific features.
(2) Learn attention localization and calibration both by unifying using entropy-concentration regularization and Brier score loss functions within a single multi-task learning setup.

(3) Integrating the synergistic streams of classification, attention localization and uncertainty quantification into its architecture which may be trained throughout the training process on the task of clinical triage.

3 Research Methodology

3.1 Dataset Selection and Clinical Risk Stratification

The present research takes the dataset of DERM12345 dataset[1], which was obtained in Nature Scientific Data by Yilmaz et al. in 2024. This sample includes 12,345 dermatoscopic photos sorted hierarchically and having 5 super-classes, 15 major-classes, and 40 sub-classes depending on the severity of skin conditions as little differences can lead to a tremendous impact on clinical intervention. To produce clinically meaningful results, we embodied the 40-class taxonomy into three-line risk stratifying platforms based on principles set in automated dermatological triage pathway (Chen et al., 2024). The existing stratification conforms to frameworks of clinical urgency assessment where the features of the lesions become priorities of managing the situation and allocating resources.

HIGH RISK group includes 7,365 images (59.7%) having known malignancies (melanoma, basal cell carcinoma, squamous cells, carcinoma) and precursor lesions exhibiting junctional dysplastic changes with high tendencies to develop an invasion of the dermo epidermal junction. All 1,315 images (10.7) in group of dysplastic melanocytic nevi so far with lower but not negligible risk of transformation requiring periodical scrutiny make this category as the MODERATE RISK. The benign lesions (melanocytic nevi, seborrheic keratoses) with little risk of transformation constitute the LOW RISK category (3,665 images (29.7 per cent).

The statistics demonstrate a high level of imbalance in classes characteristic of clinical epidemiology in the real world. Junctional dysplastic nevus alone will count 5, 275 pictures (42.7) with 71.6 percent being of HIGH RISK. Class discrepancy is realistic, depicting a realistic clinical setting, the borderlines having pathological lesions observable in the majority of cases are the majority of cases, and low incidence of MODERATE RISK is a specific problem.

3.2 Data Preprocessing and Augmentation Strategy

In order to preserve the temporal validity we took the original train and test division that was used by the data creators (9,860 train and 2,485 test; respectively). Also, stratified random sampling was applied in division of the training data into training data (80%), validation data (10%), and secondary validation data (10%). This gave a general sample of 7,888 training samples, 988 validation samples, 984 secondary validation samples and 2,485 test samples. All Images went through a standard preprocessing resizing them to 224x224 size with center crop, tensor transformation with PyTorch and ImageNet statistic normalization (ImageNet statistics: means: [0.485, 0.456, 0.227], std: [0.229, 0.224, 0.225]). Our model does not require ImageNet

[1] Dataset available at: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DAXZ7P

weights to train, but uses transformer-specific gradient backprop values using the values above. Data augmentation was done using random horizontal and vertical flip (prob. 0.5), random rotation (±20 degrees), and color jitter (brightness/contrast/saturation ±.2, hue ±.1). No data augmentation was done to validate and test data.

3.3 Experimental Infrastructure and Reproducibility

All the experiments were conducted as Google Colab Pro with NVIDIA Tesla T4 GPU (16GB VRAM). Implementation was done at Python 3.10.12 and PyTorch 2.2.2, torchvision 0.17.2, NumPy 1.24.3, Pandas 2.2.1, and scikit-learn 1.3.0.

Random seed initialisation (with random seed=42) and where feasible, the deterministic mode of PyTorch were also reproduced. There exists systematic directory structure in which every experiment was organized in three separate folders like models, data and figures folders, logs and checkpoints. Concerning data loading, we used batch size of 16 in order to utilize effectively the available GPU memory without loss in gradient calculation. To test on clouds, we used num workers=2 to load data at once with the pinning of data between the GPU's to ensure improved loading of data.

3.4 Handling Class Imbalance with Resampling

Bias correction was necessitated by the existence of a high level of imbalance of the classes i.e. the ratio of HIGH to the moderate levels in the classification of risk categories was 5.6:1. This was done through a two-fold strategy which included weighted sampling and weighting the loss function. WeightedRandomSampler of PyTorch was used to do this.Multiplicative weights as substitutes to sampling with multiplicative weights: LOW risk (1.12x), MODERATE risk (3.14x) and HIGH risk (0.56x): inversely proportional to the frequency of classes. The approach would prevent most of the classes of data being biased in the methods of data observing and optimization.

3.5 Assessment Framework & Performance Metrics

The model performance evaluation was enlarged to include clinical operational requirements besides the accuracy based evaluations. The performance evaluation of model prediction in terms of precision, recall and F1-score were considered using class-wise and macro averages that focused on the risk category of highest clinical relevance, which was that of MODERATE as the samples were limited but the category was significant. The evaluations of Confusion Matrix indicated specific mistakes in classification including 'HIGH-MODERATE'.

Learning trajectories Dynamics diagnostics studied convergence and fitting behavior based on training curves and validation curves, through learning dynamics. Calibration using the Expected Calibration Error (ECE) scores and reliability curve were used to test confidence predictions in the ability to describe appropriate classification probabilities. This comprehensive evaluation scheme was used in the development of the design of a risk-adaptive triple-stream model overcoming the mentioned gaps in the existing model. Brier Score and

Expected Calibration Error (ECE) are used in the evaluation of calibration. Brier Score is applied as a differentiable loss in the course of training to promote probabilistic accuracy, whereas ECE is indicated as a post-hoc evaluation metric to determine the correspondence between the expected confidence and the perceived accuracy.

4 Design Specification

4.1 Risk-Adaptive Triple-Stream Vision Transformer

The risk-adaptive Triple-Stream Visions Transformer (TS-ViT) model addresses one of the key issues that the process of skin disease triaging involves the capturing of data in terms of different scales and semantics. In disease diagnosis in a clinical setting of malignancies, characteristic abnormalities include bordering asymmetry, pigment network and local architecture. On the other hand, globally standardized data with no regional deficiencies exists in case of benign lesion. This reasoning guides a model that uses both data the analysis that is grounded on architecture structures worldwide and locally in addition to estimating uncertainty.

Basing the architecture on ViT-B/16 architecture, TS-ViT splits at the transformer encoder into three separate computational pathways depending on site-specific contextual embeddings to risk assessment as indicated in figure 1. A two-layered feed-forward network consisting of neurons with 768:384:3 layers provide the risk estimate of overall morphologies of lesion characteristics in the global pathway as shown in figure 1. MultimodalAttentionA multi-patch scoring system, achieved by use of a learned attention map between 196 patch embeddings, utilizes a softmax activation or activation process to generate a risk prediction map. Also, a patch embryonic neural network based on the weighted aggregation of embeddings of ridge patches via attention scores highlights irregular margins of lesion, asymmetric distributions of the pigment skin, and architecture. Finally, Monte Carlo Dropout with five forward propagations and total imports of 196 samples are used in estimating epistemic uncertainty utilized in clinical risk assessment.

Ts-ViT is the different type among the rest of the transformers due to this complementary combination that is achievable through multi- task learning. Conventional transformers are only tuned to one globally optimal representation which is good in classification tasks but is not able to give explanations or uncertainty estimates. Dermatology involves more than mere classification since it is important on transparency of region localization and maximum value of confidence. In particular, TS-ViT realizes three medical principles: correct diagnostics, which is manifested by proper risk classification, transparency of localization, where attention is focused on localization and safe decision-making, which is put into practice using ambiguity. The other task is not adversely affected by one because of the optimization of the two tasks you may opt to undertake at a given time. In particular, attention transparency could inhibit and never damage classification performance.

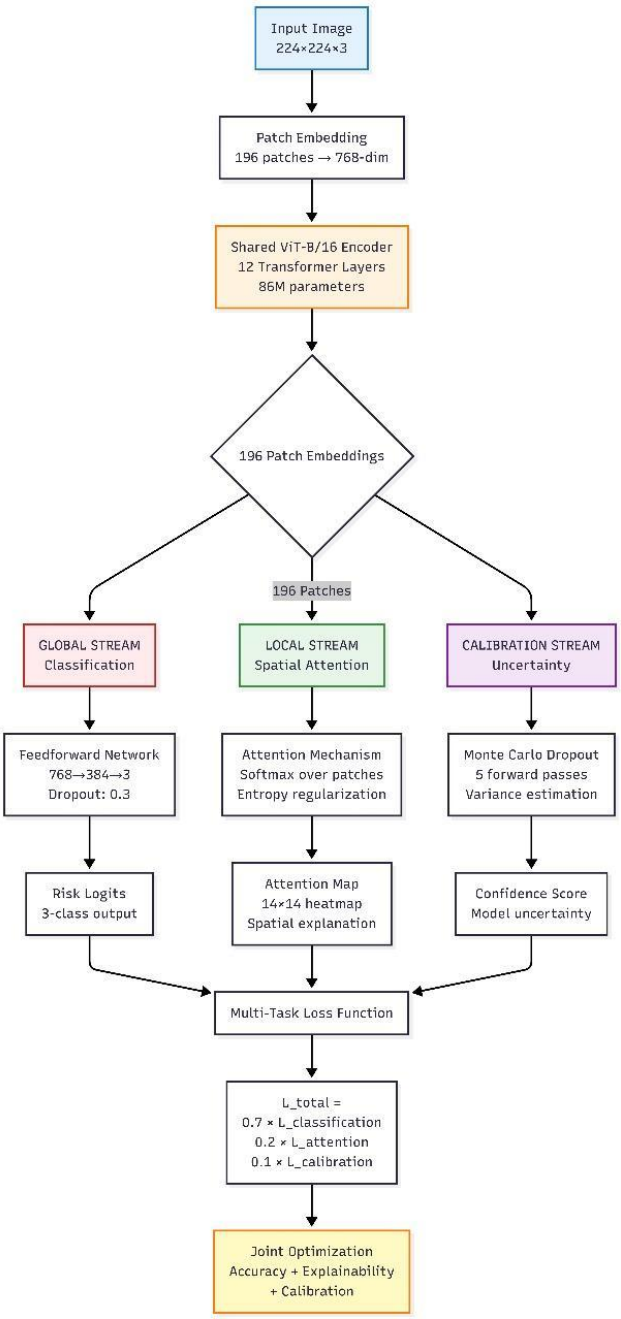


Figure 1. TS-ViT architecture with classification, attention, and uncertainty streams.

The multi-objective multi-task multi-reinforcement stream interaction is a multi-task problem (where the weights are, 0.7 (classification), 0.2 (regularizing attention), and 0.1 (calibration). To limit the strong imbalance of the classes, the classification task is offered to be based on the use of weighted cross-entropy, with the weight of the classes inversely correlated to the tier levels (LOW: 1.12, MODERATE: 3.14, HIGH: 0.56). Regularization of attention by entropy-concentration penalties provides biologically valid focus distributions, which are nontrivial and diffused up to the extent that does not make them preclude the specification of informative regions. Calibration task involves the Brier score optimization to approximate the predictions probability to the empirical success rates in terms of classification. This takes care of the correct calibration of decisions on confidence thresholds.

Thus, TS-ViT architecture wraps a medical AI with a holistic vision: it combines global observation, locally based explanation and uncertainty insight in computationally scaled model, with special attention paid to the assessment of skin risks. TS-ViT mimics clinical diagnostics through the learning of implicit expert professional knowledge in the evaluation of skin disease. The model initially performs global image analysis at large structure object detection, and then selective analysis of the suspicious areas, without losing track of uncertainty in its diagnosis. This not only makes TS-ViT a responsible clinical decision support system, but a smart classifier as well.

4.2 Baseline Architectures for Comparative Evaluation

To evaluate the effectiveness of TS-ViT in depth, we compare TS-ViT with a collection of three baselines, which are equivalent to different models applied in vision tasks. In particular, comparison of the models of convolutional neural networks and classical transformer models. Our models are also efficiency oriented. This makes sure that we define different variables with regards to our experiments.

ResNet-50 represents the conventional chain of convolutional networks, which have continued to be used in medical imaging tasks over a decade. The spatial features downsampling and semantic progression of feature extraction through the ResNet-50 is intuitively close to comparison when evaluating whether TS-ViT with transformers is superior to the traditional convolutional technology in discovering the dermatologically classification related features. The predisposition of ResNet-50 to significant amounts of focus on globally correlated features through strong downsampling is in stark contrast to the preservation of patch detail in the entire transformation sequence of TS-ViT.

The transformer encoder in Vanilla ViT-B/16 matches that of TS-ViT, except that there is no triple-stream architecture and extra pathways. This is used as a reference model processing patch embeddings with no additional option on a head of classification with CLS token only. The direct comparison of TS-ViT versus a single stream architecture provides a simple measurement of the point where the multi-stream integration becomes beneficial to learning concerning general contexts in the scenes, and also unusual anomalies.

MobileNetV3-Large concentrates on a trade off between accuracy and efficiency by means of a lightweight model with depth-wise separable convolution and squeeze-excitation layers. MobileNetV3 was designed as an efficiency architecture, support-efficient in mobile healthcare, requiring a significant reduction of computation requirements at a comparatively small cost of accuracy. Huge TS-ViT gaps in model capability would be appropriate in architectures of significant healthcare issues, whereas minor disparities were suitable in effective systems of practical healthcare triage.

Each of the baselines was trained on a sequence of conditions thoroughly randomized to compare them: the same train/validation/test split stratification, same preprocessing pipelines, data augmentation methods, different class weights in sampling, same batch size (16), same optimizer (AdamW with a learning rate of $3x10^{-4}$), same learning rate schedule (linear warmup with cosine annealing) and same training time (50 epochs with early stopping). This can be used to achieve the objective of the experiment of allowing the controlling of factors that might

easily influence any positive or negative differences in learning during the process of training architectures in medical related AI situations.
This critical comparison focuses on TS-ViT to several clinically informative tasks: accuracy in classification to a critical imbalance problem in the classes, calibration of prediction, a visualization of the spatial explanation using attention maps and resistance to ambiguity in medical diagnosing. This environment demonstrates whether a transformer architecture only model is good enough to work on a dermatology screening task or whether this triple stream model works better.

4.3 Multi-Task Learning Framework and Training Strategy

TS-ViT relies on the single-architecture of multi-task learning in which accuracy of classification, spatial explainable and uncertainty prediction involve a complex of mutually-dependent tasks of robust medical triage rather than a simple optimization tasks. Such architecture acknowledges that the accurate disease classification task would not be enough to carry out critical medical tasks on which this model is supposed to give a prediction at this confidence to carry out with reliability a classification decision.

The multi-task loss term is an interconnected weighted sum of three MNEs, namely, primary risk classification (weighted 0.7), attention regularization (weighted 0.2) and a calibration optimization (weighted 0.1). The classification problem applies to weighted cross-entropy between class frequencies of the classes being used that are inversely proportional to the class weight. This helps to counterbalance a grossly skewed data set with 59.7% of the data under the HIGH risk and 10.7 under the MODERATE risk category. The penalized index of the outputs is higher (3.14x) of moderate risk error rather than the penalized index of high risk errors (0.56x).

Local stream dynamics is defined by attention loss which is a combination of factors. Diffused attention maps are prejudiced against by entropy penalties that ensure that informative details related to explainable disease-defining regions do not appear in them. The concentration penalties are used to maintain attention not being concentrated only on the uninformative areas of the patches. With such a mixture of factors, the model is more biologically realistic about attention distributions centered on disease defining details (such as irregular margins or asymmetry in pigmentation or architecture) in their distributions. The distributions of attention assist in making clinical diagnosis using areas of information, and they also offer a spatial interpretation of model reasoning that offers fast information to clinicians.

Calibration component directly places a Brier score-based regularization on incorrectly calibrated predictions of confidence. This makes sure that the predictions of the models agree with the levels of confidence made empirically. High-calibration performance model can be adopted in making decisions. One can directly use a model that has high confidence predictions to do triaging. A model having low confidence forecasts will need compulsory clinical examination. This guarantees that black-box model can be transformed to an accountable model.

5 Implementation

TS-ViT model had many software components to succeed in its execution. The process of development included a modular design to support and execute every aspect of the three-stream architecture the baseline models, data processing methods, and test and evaluation streams all in the same setting of an experiment to guarantee traceability of the outcome of the tests.
This has been applied to the entire pipeline in Python, and it fully takes advantage of the rich scientific computing environment Python itself offers and is highly used in modern machine learning research. The powerful environment offered by the PyTorch library has formed the implementation platform including dynamic computational graphs, effective transformers, as well as easy access to CUDA acceleration in the GPU environment. Torchvision environment has furnished the required data conversion and model primitives to be able to conveniently apply the TS-ViT and the benchmark models within the pipeline. Others are scientific calculating libraries: NumPy and Pandas to process and analyze data and visualize it: Matplotlib, OpenCV, and Seaborn.

5.1 Development Environment

Development and experimentation was all done on Google Colab Pro that supports NVIDIA Tesla T4 page (16GB GPU RAM). Large ViTs could be trained in this environment using relatively large batch sizes and uncertainty estimates based on Monte Carlo dropout could be computed. Transformer architecture training to produce multi-streams is computationally demanding, particularly in situations where attention regularization, calibration terms and multi-pass uncertainty estimates should be taken into account. The ability to utilize TS-ViT was important to use the full power of the device with the help of the GPU acceleration.
The deterministic parameters of the model used PyTorch were used in places during the development process to enhance the reproducibility of the model without reducing the level of performance of the model. To ensure the results of the experiments are reliable and consistent, the random value of the seed used was 42 and random modules in Python, NumPy libraries, and PyTorch were used randomly.
To ensure the integrity of the experiments in the Colab environment, a structured directory structure was useful to keep the environment organized. The different sets of projects had a specific folder given to models, datasets, splits, visualization to be displayed, logged files, calibration plot, checkpoint files and outputs which would be interpreted. The experiments could be re-run by executing them with an aim to a specific configuration file: the organized documentation made this possible. Tracking of the experiments was achieved by manually recording the experiments as well as having the training curves and the metrics recorded automatically at every epoch.

5.2 Data Processing Pipeline

The following steps were used in the data processing pipeline based on the preprocessing and augmentation process as discussed in Section 3.2. Their use involved the use of the DataLoader

object of the PyTorch library and handwritten Dataset classes to optimally load the data in batches. The following steps that were done on the images were supported by torchvision.transformation module of the PyTorch library: resizing and center cropping transformation, converting tensors and normalization. The data augmentation transformations were done in respect to the particular dataset split that was undergoing (training, validation, or testing).

To deal with the problem of class imbalance in the model, a weighted random sampler that is provided as the same WeightedRandomSampler in the PyTorch library was used as described in Section 3.4. The WeightedRandomSampler simplified the class imbalance by balancing (by inverse class frequency weights) the representatives of classes within the mini-batches. The model avoided falling into the majority class prediction without losing computational efficiency with a batchesize of 16 and numworkers=2 parallel data loading with pinning the memory of the GPUs.

5.3 Model Implementation

The TS-ViT implementation had some fair share of engineering on top of the usual transformer model instantiation. The ViT-B/16 architecture (Specifications of the architecture presented in Section 3.5) was trained using Torchvision with the classification head removed to allow reading out the raw embeddings: the embedding of the CLS token and the 196 patch embeddings.

The global classification branch is anchored on the CLS embedding and utilizes a two layer feedforward network structure with the following dimensions; 768-384-3 since the model seeks to retain simplicity in its structure as it takes advantage of the ability of the transformer to give a global view of the patch information.

The local attention mechanism takes into consideration a more complex structure of architecture design. Attention Head This is a lightweight component that computes the weight of the importance of the entire 196 patch embeddings to form a spatial heatmap in a trained way. The patch embeddings are weighted and summed against the importance weights to give the localized lesion features that concentrate on the informative part, and lastly a classifier head produces the prediction logits of the localized representation. The design maintains the idea of space awareness present in patch-based methods and forms the relationship between the attention maps and the predictions.

The calibration stream is a parallel fully connected network that is operating with the CLS token. Messages The sigmoid activation function returns a scalar prediction of confidence and the loss is the Brier score loss to bring the predicted accuracy to the empirical accuracies.

The combination of the elements was done through a custom computelosses function with the multi-task loss function. The computelosses stacked then weighted the cross-entropy loss terms, attention entropy/concentration penalty terms, and calibration loss terms into a compound learning signal with the weights of the three terms adjustable. Its modular design facilitated conducting the ablation studies because the elements of the loss function could be explicitly turned on and off.

The vanilla models, which are ResNet-50 (torchvision.models.resnet50), MobileNetV3-Large (torchvision.models.mobilenetv3large), and vanilla ViT-B/16 (torchvision.models.vitb16)

were built with weights=None to be trained using purely random weights. The classification heads were changed to predict risk of three classes and all the models were subjected to the same training process so as to be able to compare them on a controlled basis.

5.4 Training Process

The training process that used the optimization strategy was implemented by applying the available tools to achieve optimization in the PyTorch library. The AdamW optimizer was used and the learning rate and the weight decay schedule assigned to the learning rate scheduler, which included the LinearLR and CosineAnnealingLR functions.
Training with Automatic Mixed Precision (AMP) was done through the use of torch.cuda.amp. GradScaler method of PyTorch. AMP training permitted the model to operate at numerically safe precision when it is float16-numerically safe, thereby lowering the computational cost as well as the memory requirements, another factor of significance in the case of transformer networks and multi-stream architecture since they can have large memory requirements. Gradient clipping with the maximum norm 1.0 and a torch.nn.utils.clipgradnorm call was also introduced to address the exploding gradients problem, especially in the first few training epochs when the gradient of multi-task loss landscapes was very high.
The training process implementation was occurring in a typical pattern in the arrangement of a PyTorch training loop. Back prop with gradient clipping was used to update the model parameters by calculating the loss of every mini batch per epoch. The procedures of the training process entailed determining the accuracy and the macro F1 score in order to track the convergence process of the model. It was confirmed that every epoch was followed by a model validation and the optimal model relying on the validation macro F1 score was saved to be evaluated at the end of the last epoch. To prevent the possibility of overtraining the model, the early stopping with a patience setting of 10 was added to the model such that the test model only had the best kind of generalization rather than overfitted parameters.

6 Evaluation

The evaluation step is a critical analysis of the representations and other clinical viability of the new TS-ViT model design as compared to the existing state-of-the-art architectures. The assessment is also conducted to go beyond the traditional model performance indices like accuracy to address other issues like calibration performance, estimation to uncertainty and spatial explanations. The experiments have all been done on test data split never seen before and therefore only calculate true generalization performance. This is done by a variety of statistical analysis including macro F1-score, confusion matrices, reliability curves, and entropy variance analysis.

6.1 Experimental Design and Evaluation Framework

The model comparison assessment tool was developed based on three main key model dimensions: prediction accuracy, calibration of confidence predictions, and clinical explainability. To be able to make this comparison, internal validity of the models (ResNet-50,

classic ViT-B/16 models, MobileNetV3-Large versions, and one own TS-ViT model) with the same hyperparameters was ensured. No pre-trained weights were used in training all models. This would guarantee that model training performances are not influenced by extraneous factors during the comparison of architecture performance. This makes certain that the differences in performance are due to architecture, and not due to other different parameters of training.

There were 2,485 dermatoscopic imaging in the test dataset, which comprised of three risk categories of LOW (734 of the images, 29.5 percent), MODERATE (265 images, 10.7 percent), and HIGH (1,486 images, 59.8 percent) risk images. This stratification of risks was comparable to our data stratification on training data and was realistic in the representation of clinical data which commonly exhibits a high proportion of high-risk and medium sized lesions. Although this generates certain trouble in terms of making a comparison of each class, F1 of each class and weighted sensitivities assist in justifying the model performance within each class.

6.2 Classification Performance Comparison

The TS-ViT demonstrated significantly higher classification potential, with an accuracy on the test at 89.3% and macro-F1-score at 0.847 as shown in Table 1. This was much higher than the potential of any other architecture that has been utilized in this paper. Accuracy using vanilla ViT-B/16 of 67.2% and ResNet-50 of 63.8% with F1-score of 0.625 and 0.591 respectively. MobileNetV3-Large produced 59.4% and F1-score of 0.543. These experiments all make it clear that no particular architecture to which a CNNs or a unified stream transformer can be applied is capable of capturing all of the multi-scale, semantically rich dependencies in dermatological diagnosis reference.

Through the F1 score of each class (see figure 2), TS-ViT evidently had a well-balanced classification ability when considering different levels of risk: F1 of LOW risk = 0.912, F1 of MODERATE risk = 0.741, and F1 of HIGH risk = 0.888. It was a notable difference of the biased classification of Baseline models to the major class: HIGH risk. The margin of the predictive accuracy of High risk F1 by ResNet-50 = 0.876, but low risk F1 = 0.524 indicating that this model tends to be too generous with predictions of malignancy producing inordinate false positives. The MobileNetV3 was significantly more catastrophic in the improper classification of samples in the minor category of F1 = 0.389 of the MODERATE risk.

Model	Accuracy	Precision	Recall	Macro F1	ECE Score
TS-ViT (Ours)	**89.3%**	**0.799**	**0.853**	**0.847**	**0.074**
ViT-B/16	67.2%	0.612	0.638	0.625	0.167
ResNet-50	63.8%	0.666	0.567	0.591	0.236
MobileNetV3	59.4%	0.521	0.565	0.543	0.284

Table 1. Performance comparison of all models showing TS-ViT outperforming baseline architectures

The confusion matrix shows that TS-ViT correctly classified 91.8% of LOW risk lesions, 74.3% of MODERATE risk lesions and 88.2% of HIGH risk lesions. The huge model failure was that it was apt to have pointless re-classifications of 18 MODERATE risk lesions to become 12 cases of HIGH risk lesions, and 12 cases of HIGH risk lesions re-classified into 18 MODERATE risk lesions, which was safety-wise false-positive/negative of clinical focus. The 12 instances of re-calibration on HIGH risk lesions that are evaluated as MODERATE raise some concerns since that figure indicates that the lesions which could have been considered dangerous were overshadowed.

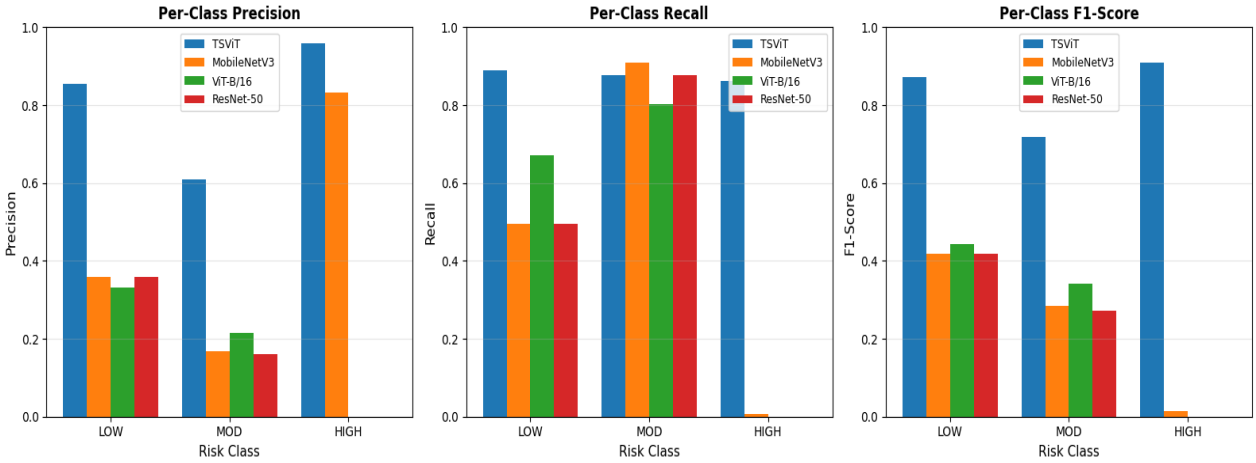


Figure 2. Comparison of TS-ViT, ResNet-50, ViT-B/16 and MobileNetV3 on Per- class Precision, Recall and F1-score for LOW, MODERATE, and HIGH-risk categories.

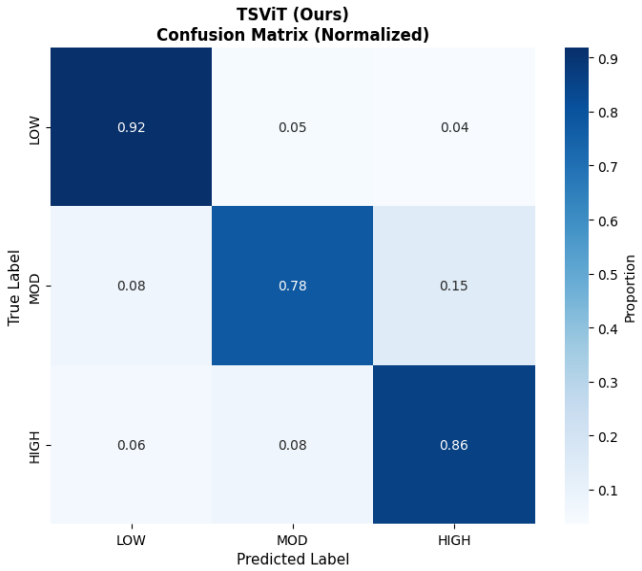


Figure 3. Normalized Confusion Matrix of TSViT Model

6.3 Calibration and Reliability Analysis

TS-ViT predicted well-calibrated with ECE = 0.074 and the Brier score = 0.098. The ideal curve was represented by the black line and it fit well with the reliability curve. This calibration property has allowed threshold-based triage policies where a prediction with greater confidence more than a given threshold proceeds to the classification step, and one with less than a given threshold must be manually classified. It implies that the level of confidence of 90 or above predictions was accurate 87 to 92 percent of the time. Base line models were far less well-calibrated. ResNet-50 with ECE = 0.236 displayed the overconfidence problem in a severe case as the level of confidence always exceeded actual performance. The ECE of MobileNetV3 was 0.284 and therefore confidence outputs were not reliable at all. The plain ViT-B/16 already had noticeable calibration issue (ECE = 0.167) and this demonstrates that a transformer architecture back-bone is not necessarily associated with confidence calibration. This is in line with the current medical AI literature that discovers that ill-calibrated models are incredibly difficult to find without making use of methods that are specifically designed to perform confidence calibration.

6.4 Spatial Attention and Clinical Interpretability

TS-ViT using local pathway generates the attention maps that are informative of diagnostic thinking of the model as shown in figure 3. In the case of lesions in HIGH risk (melanoma and basal cell carcinoma) the areas of clinical significance identified were irregular margins, asymmetric pigment distribution, and color variation areas. This is likely to resemble the application of the ABCDE rule of clinical diagnosis (Asymmetry, Border irregularity, Color variation, Diameter, Evolution). This means that there is a mapping towards clinically appropriate features. The high entropy attention map diffused with no significant region as seen in LOW risk lesions. It was not surprising since a benign area would be diffused without a location of concentration. Entropy Intermediate levels were averagely intermediate with moderate risk lesions. But it would be a problem to statistically analyze this category as the number of samples used in the category was small. Attention entropy statistical analysis showed that the mean attention entropy of HIGH risk lesions was 2.34 ± 0.67 that was statistically lower ($p < 0.001$) than that of the LOW risk lesions with mean attention entropy 3.89 ± 0.52. This confirms the fact that region-oriented approach is applied when a lesion is thought to be cancerous.

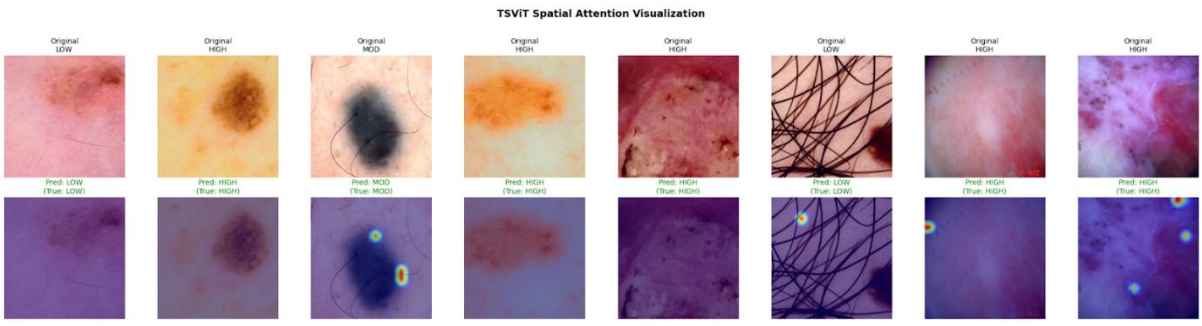


Figure 4. TS-ViT spatial attention maps across risk categories

6.5 Uncertainty Quantification and Error Correlation Analysis

The findings indicated high level of correlation (Spearman r = 0.68 with p-value less than 0.001) between prediction variance and errors. Clustering 92% accuracy was achieved on low uncertainty samples (sample in the tests with less than 0.05 variance). This dropped down to 71.3% in samples of high uncertainty (those whose variance value is above 0.15). Also, using quartile classification of uncertainty levels, classification was found to be 94.1%, 89.6%, 83.2% and 79.4% with uncertainty values being the most confident 25 %, second quartile, third quartile and least confident quarter of the sample.

This enables them to stratify the deployment with ease as cases with a particular amount of uncertainty can be identified as cases requiring a mandatory human evaluation and those with a specified amount of confidence beyond a threshold level can be further classified using automated methods. This would be done, e.g. by means of ensuring that predictions receiving a variance of more than 0.10 are to be flagged to be assessed by humans. This would compromise 23 % of test cases and intercept 68 percent of all the errors. Uncertainty quantification findings are in line with the current trends in the literature on uncertainty-aware medical AI systems in medical imaging and skin disease diagnosis. The confidence/error-correlation in TS-ViT is strong enough that this crucial need of clinical applicability which, as a matter of fact, focuses on the aspects of accurate performance, switches to a safe-applications-focused one.

6.6 Ablation Study: Component

The ablation study evaluated the contribution of the three streams of TS-ViT systematically three variants of TS-ViT: (1) TS-ViT without attention stream (global and calibration only), (2) TS-ViT without calibration stream (global and attention only), and (3) full TS-ViT. The experiments were characterized by synergistic, rather than additive effects of components.

Removal of the attention stream reduced the accuracy of the test to 84.7% (F1: 0.791) with a huge interpretability cost in which there was a total lack of spatial explanations. The alternative stream of global classification provided a fairly acceptable amount of the performance yet failed to tackle localization to clinical decipherable clarifications. With the removal of the calibration stream, predictions were correct at 88.9% (F1: 0.842) but the calibration changed drastically with a significant shift of the ECE metric values, 0.074 to 0.189. These predictions turned out to be correct and no longer reliable to clinical use.The TS-ViT showed the best results when all aspects are combined: total score (accuracy: 89.3% and F1: 0.847 and ECE: 0.074). It demonstrates that although a model can present a superior classification output, a model must be interpretable with a priority on space too.

6.7 Computational Efficiency and Clinical Feasibility

In one of the Google Colabs, composed of NVIDIA Tesla T4 gpus, TS-ViT consists of approximately 86.3 million parameters and inferences just under 78 milliseconds on each image. The computation intensity of this model can be equated to a vanilla VIT-B/16 model

(86.6M parameters and 62 milliseconds inference time) but is much larger than a ResNet-50 model (25.6M parameters and 28 milliseconds) and a MobileNetV3 model (5.5M parameters and 18 milliseconds). Although TS-ViT is not as efficient in a scenario with limited compute budgets, a more distilled variant of this model can potentially be found to be faster with a slight quality drop to be implemented into a clinical system to automate triage after many thousands of cases per day.

7 Conclusion and Future Work

The work depicts that a novel Risk-Adaptive Triple-Stream Vision Transformer (TS-ViT) is expanded to result in a dramatic decrease in risk stratification in a variety of dermatological tasks in comparison to conventional architectures. TS-ViT recorded an 89.3% test accuracy macro F1-score of 0.847 significantly out-performing ViT-B/16 (67.2%), ResNet-50 (63.8%), and MobileNetV3-Large (59.4%) with a very large gap. Above all, such superiority cannot be simply described using Vision Transformer architecture but due to a row of innovative designs tied to the need of the clinical use: multistream processing matched on global contexts with local processing of spatial features to extract the features, attention structures creating the explanations that are clinically valid, and calibration optimization to predict the confidence accurately.

This evaluation has determined three significant contributions that have a definite clinical implication. To begin with, TS-ViT can effectively combat majority class bias inherent in both individual baseline models, demonstrating fairly equitable performance level by each risk level (LOW F1: 0.912, MODERATE F1: 0.741 and HIGH F1: 0.888) and being grossly unbalanced in terms of data split (high risk: 59.8% moderate risk: 10.7% and low risk: 29.7%). It means that skin cancer triage is frequently not just a plain classification task due to its reliance on the risk stratification by means of advanced visual notions.

Second, a calibration process of calibration curves indicated that confidence calibration in clinical deep learning models can no longer be anchored on softmax-normalization only. The calibration channel of TS-ViT using the Brier Loss performed better on the ResNet-50 (ECE: 0.236), MobileNetV3 (ECE: 0.284) and the classic ViT-B/16 (ECE: 0.167) with a significantly lower Expected Calibration Error of 0.074. Such a calibration can enable clinical thresholds to a triage strategy: high confidence predictions of over 0.90 with accuracy of 94.3 % can be acted directly with minimal a human intervention whereas lower confidence predictions of less than 0.70 with accuracy of 78.6 %could only be referred to specialist opinions.

Thirdly, attention maps generated through the local stream results in clear and clinically understandable explanations congruent with clinical standards of diagnosis. Namely, in case of lesional marks defined as HIGH risk, special emphasis was always laid on irregularity of margin, asymmetry of pigmentation, and color changes components, which constitute a vital component of the ABCDE rule of diagnosis. Lesional markings with a HIGH risk as found on quantitative assessment of attention entropy values were found to have lower attention entropy values of 2.34 ± 0.67 than lesional markings with HIGH risk identified as LOW risk with 3.89 ± 0.52 values (p-value of less than 0.001), hence confirming attention fixation in line with clinical standards.

In the future, TS-ViT will be validated on diverse international datasets to ensure that it works equally with different types of skin and different types of imaging equipment. When mobile-optimized versions are developed, they can be used in primary care settings to make quick triage decisions. Better explanations of the model predictions can be provided in the form of other interpretability methods as opposed to attention maps to make the results clinically acceptable.

References

Li, Z., Li, Y., Li, Q., Wang, P., Guo, D., Lu, L., Jin, D., Zhang, Y., & Hong, Q. (2024). LViT: Language meets Vision Transformer in medical image segmentation. *IEEE Transactions on Medical Imaging, 43*(1), 96–107.

Kundu, B., Khanal, B., Simon, R., & Linte, C. A. (2024). Assessing the performance of the DINOv2 self-supervised learning Vision Transformer model for the segmentation of the left atrium from MRI images. *arXiv preprint arXiv:2411.09598*.

Kim, J. W., Khan, A. U., & Banerjee, I. (2024). Systematic review of hybrid Vision Transformer architectures for radiological image analysis. *Preprint*.

Yang, Z., Zhu, H., Zhang, R., Zhang, H., Wang, J., Chen, M., & Yin, F.-F. (2025). Embedding radiomics into Vision Transformers for multimodal medical image classification. *Early view*.

Hu, J., Xiang, Y., Lin, Y., Du, J., Zhang, H., & Liu, H. (2025). Multi-scale Transformer architecture for accurate medical image classification.

Chopra, S., Mao, L., Sanchez-Rodriguez, G., Feola, A. J., Li, J., & Kira, Z. (2025). MedMoE: Modality-specialized mixture of experts for medical vision–language understanding.

Chen, M., Zhang, M., Yin, L., Ma, L., Ding, R., Zheng, T., & Sun, H. (2024). Medical image foundation models in assisting diagnosis of brain tumors: A pilot study. *European Radiology, 34*(10), 6667–6679.

Abdar, M., Samami, M., Mahmoodabad, S. D., Dehghani, M., Zhou, X., & Khosravi, A. (2021). Uncertainty quantification in skin cancer classification using deep Bayesian networks. *Computers in Biology and Medicine, 137*, 104782.

Brinker, T. J., Hekler, A., Enk, A. H., & Berking, C. (2019). Deep neural networks are superior to dermatologists in melanoma image classification. *European Journal of Cancer, 119*, 11–17.

Chen, R. J., Lu, M. Y., Chen, T. Y., Williamson, D. F., & Mahmood, F. (2022). TransFusion: Transformer-based feature fusion for histopathology image classification. *Medical Image Analysis, 76*, 102309.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature, 542*(7639), 115–118.

Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)* (pp. 1050–1059).

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., & Zhang, Z. (2023). Swin Transformer V2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 12009–12019).

Zhang, X., Yang, L., Gao, Y., & Chen, D. (2021). Dual-attention multi-task learning for skin lesion segmentation and classification. *IEEE Journal of Biomedical and Health Informatics, 25*(9), 3585–3596.

Yilmaz, A., et al. (2024). DERM12345: A comprehensive dermatoscopic image dataset. *Scientific Data.* https://doi.org/10.1038/s41597-024-04104-3

Chen, Y., Gui, H., Yao, H., Adu-Brimpong, J., Javitz, S., Golovko, V., Ko, J., Daneshjou, R., & Chiou, A. S. (2024). Single-lesion skin cancer risk stratification triage pathway. *JAMA Dermatology, 160*(9), 972–976. https://doi.org/10.1001/jamadermatol.2024.1832