XAI-Driven Skin Cancer Diagnosis

Kunwar Ranjeet
Student, CINTEL
SRM University
Chennai, India
Ranjeetkunwar2018@gmail.com

Subham Saha
Student, CINTEL
SRM University
Chennai, India
Subhamsaha132003@gmail.com

M. Ratnakumari
Asst. Professor, CINTEL
SRM University
Chennai, India
madukurr@srmist.edu.in

***Abstract*—** Skin cancer is one of the most prevalent and life-threatening diseases worldwide, emphasizing the need for accurate and early diagnosis to improve patient outcomes. This research presents an advanced skin cancer detection frame- work that integrates the Swin Transformer (Shifted Window Transformer) with semantic segmentation techniques for the classification of dermatoscopic images. The Swin Transformer effectively captures hierarchical global features, enabling precise differentiation between melanoma and benign lesions. To enhance the model's generalizability, preprocessing techniques such as image resizing, normalization, and augmentation are applied to datasets including HAM10000, ISIC-2008, ISIC-2019, and PH-2. Furthermore, Explainable AI (XAI) methodologies, including Grad-CAM and Grad-CAM++, are incorporated to provide visual explanations, highlighting crucial regions influencing classification decisions. The model's performance is evaluated using accuracy, sensitivity, specificity, and the Dice similarity coefficient, ensuring reliable and interpretable diagnostics. Additionally, Bayesian Optimization is employed for hyperparameter tuning, optimizing model performance while maintaining computational efficiency. This research contributes to the field of AI-driven dermatology by offering a transparent and robust diagnostic tool that assists dermatologists in early melanoma detection. Future enhancements include optimizing the model for real-time edge deployment, expanding datasets to improve generalizability across diverse skin tones, and integrating multi-modal data for enhanced diagnostic accuracy.

Index Terms - Interpretable Machine Learning, Swin Trans-former, Explainable AI, Skin Lesion Classification, Attention Mechanism, Model Transparency, Performance Metrics.

I. INTRODUCTION

Skin cancer is one of the most commonly diagnosed cancers worldwide, with millions of new cases reported annually. Early and accurate detection is crucial for improving patient outcomes, as timely intervention significantly increases the chances of successful treatment. Traditional diagnostic methods rely on dermatoscopic examination by medical professionals, which, despite being effective, is often time-consuming, subjective, and dependent on clinical expertise. In recent years, artificial intelligence (AI) and deep learning models have demonstrated remarkable potential in automating and enhancing skin cancer detection, providing reliable and efficient diagnostic solutions.

This project introduces an advanced AI-driven skin cancer recognition system leveraging the Swin Transformer (Shifted Window Transformer), a state-of-the-art deep learning model that captures hierarchical and global contextual features for precise lesion classification and segmentation. The system is designed to analyze dermatoscopic images from datasets such as HAM10000, ISIC-2008, ISIC-2019, and PH-2, ensuring robust and generalizable predictions. Explainable AI (XAI) techniques, including Grad-CAM and Grad-CAM++, are integrated to improve interpretability, allowing medical professionals to understand the reasoning behind the model's predictions.

To further enhance performance, Bayesian Optimization is employed for hyperparameter tuning, ensuring optimal accu- racy and computational efficiency. The system is evaluated using accuracy, sensitivity, specificity, and the Dice similarity coefficient, validating its reliability in real-world applications. By combining deep learning with explainability, this project aims to bridge the gap between AI-driven diagnostics and clinical adoption, providing dermatologists with a transparent, accurate, and efficient tool for early melanoma detection. Future developments will focus on real-time deployment on edge devices, dataset expansion for diverse skin tones, and multi-modal integration to further refine diagnostic precision.

A. Transformer Model

The Vision Transformer (ViT) and Swin Transformer are two advanced deep learning architectures designed for computer vision tasks, particularly image classification and object detection. ViT, introduced by Google, treats an image as a sequence of non-overlapping patches, similar to tokens in NLP, and processes them using self-attention mechanisms. This allows ViT to capture global dependencies across an image efficiently. However, ViT lacks spatial hierarchy, making it less effective for fine-grained image understanding and computationally expensive when processing high-resolution images.

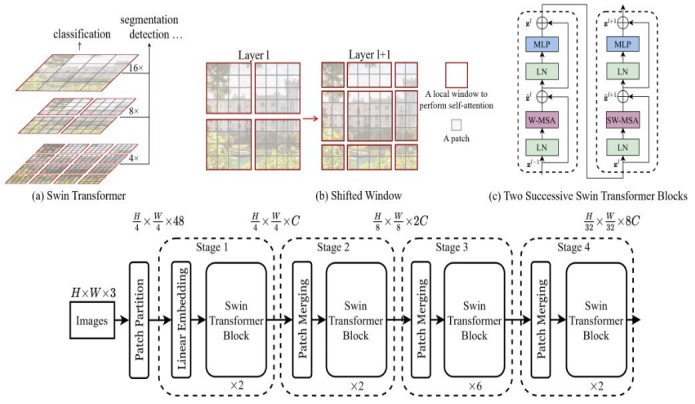

Fig. 1. Transformer Model

On the other hand, the Swin Transformer (Shifted Window Transformer), introduced by Microsoft, builds on ViT's concept but improves efficiency by introducing a hierarchical architecture with shifted window-based self-attention. Instead of processing the entire image globally like ViT, Swin Transformer divides the image into local windows, allowing it to maintain spatial locality while progressively expanding the receptive field. This hierarchical design significantly reduces computational complexity, making Swin Transformer more scalable and efficient for tasks like image segmentation, object detection, and medical imaging. Unlike ViT, Swin's ability to capture both local and global contextual information makes it particularly effective for skin cancer detection, where fine details in dermatoscopic images are crucial for classification.

B. *Explainable Artificial Intelligence*

Explainability and interpretability in deep learning (DL) models are rapidly evolving areas in medical AI research, particularly for skin cancer detection. Explainable AI (XAI) techniques play a crucial role in enhancing transparency by highlighting the key regions in dermatoscopic images that contribute to the model's classification decision. By providing visual and feature-based insights, XAI improves trust, user acceptance, and clinical adoption of AI-based diagnostic tools. Furthermore, it helps in identifying biases in the model, ensuring fairness across diverse skin tones and lesion types. When a model produces unexpected results, analyzing its decision-making process can help in debugging and refining the system to enhance its reliability. Explainability also aids in selecting the most suitable deep learning model for specific diagnostic applications, balancing complexity, accuracy, and interpretability to ensure effective deployment in healthcare environments.

Several XAI techniques have been explored in the literature for medical imaging tasks. SHAP (Shapley Additive Explanations) analyzes feature importance scores to determine the most influential visual or textual inputs affecting predictions. LIME (Local Interpretable Model-Agnostic Explanations) breaks down an image into interpretable components, perturbs the data, and fits a simpler model to explain how variations influence predictions. Layer-wise Relevance Propagation (LRP) assigns importance scores to neurons within the neural network, helping to visualize critical decision- making paths. Saliency maps, such as Grad-CAM and Grad- CAM++, generate heatmaps to highlight significant regions in dermatoscopic images that influence classification out- comes. These techniques provide visual explanations, improving model transparency and aiding dermatologists in clinical validation. Additionally, deep learning architectures with built- in interpretability, such as attention mechanisms in transformers, reveal which parts of an image the model focuses on for classification. Integrated Gradients compute the importance of input features by tracking changes in gradients from a baseline input to the actual image, further enhancing model interpretability.

Evaluating XAI representations requires multiple assessment metrics based on model performance and application context. User studies involving dermatologists help measure how informative and trustworthy AI explanations are in real- world scenarios. Qualitative metrics assess the faithfulness and consistency of model explanations by comparing them against ground truth annotations or tracking how they change with slight variations in input. Moreover, quantitative metrics such as coverage, precision, and recall ensure the completeness and accuracy of AI-generated explanations. A combination of these evaluation techniques ensures that explainability is robust, reliable, and clinically meaningful, making deep learning solutions in skin cancer detection more interpretable, transparent, and actionable for medical professionals.

II. Literature Survey

A. Related Work

With an emphasis on datasets like PH2, ISIC, DERMOFIT, and MEDNODE, numerous studies have investigated machine learning and deep learning approaches for the categorization of skin lesions.

In order to facilitate medical diagnosis, Ozkan and Koklu (2017) used machine learning models to preclassify skin lesions into normal, abnormal, and malignant categories. Four machine learning (ML) techniques were examined in their study using dermoscopic images from the PH2 dataset: artificial neural network (ANN), support vector machine (SVM), K-nearest neighbor (KNN), and decision tree. The corresponding accuracies were 92.50 percent, 89.50 percent, 82.00 percent, and 90.00 percent.

Using the PH2 dataset, Alkarakatly et al. (2020) created a five-layer convolutional neural network (CNN) to categories skin lesions as either nevus or melanoma. The model's overall accuracy was 95 percent, with 94 percent sensitivity, 97 percent specificity, and 100 percent AUC.

Using the DERMOFIT and MEDNODE datasets, Mukher jee et al. (2019) presented CNN Malignant Lesion Detection (CMLD), a CNN-based method. With DERMOFIT, their model achieved 90.58 percent accuracy, with MEDNODE, 90.14 percent, and with both datasets together, 83.07 percent. Using the ISIC dataset, Shahsavari et al. (2022) created the Ensemble of Deep (SLDED) model for skin lesion identification. They achieved a mean average precision (mAP) of 0.96 by utilizing a VGGNet feature extractor in conjunction with a modified Faster R-CNN. With high precision (87.1 percent and 90.2 percent) and AUC scores (98.6 percent and 98.1 percent), their classification studies on the ISIC and PH2 test sets yielded accuracy rates of 97.1 percent and 96 percent, respectively.

For automatic skin lesion segmentation, Jiang et al. (2020) suggested the Channel and Spatial Attention Residual Module (CSARM-CNN) system. Their model produced accuracy values of 94.96 percent and 95.23 percent, as well as specificity scores of 99.03 percent and 99.45 percent, using the ISIC 2017 and PH2 datasets.

Using the ISIC dataset, Kumar and Vatsa (2022) examined and evaluated decision tree-based algorithms (XG-Boost) and deep neural networks. They used metrics for loss, precision, recall,

ROC, and F1 score to assess classification performance. Bidirectional RNN surpassed other RNN architectures with an accuracy of 95.96 percent, whereas CNN's VGG16 architecture performed best with an accuracy of 89.6 percent. The accuracy rate of the XG-Boost method was 97.22 percent.

In order to categories three distinct kinds of skin lesions, Hosny et al. (2018) suggested a deep transfer learning method that modified AlexNet with a SoftMax layer. Their model performed well with 98.61 percent accuracy, 98.33 percent sensitivity, 98.93 percent specificity, and 97.73 percent precision using the PH2 dataset.

An interpretable ensemble stacking method for melanoma detection was introduced by Alfi et al. in 2022. They integrated deep learning architectures (MobileNet, ResNet50, Xception, DenseNet121, and ResNet50V2) with machine learning models (ML) (logistic regression, random forest, SVM, XG-Boost, and KNN). Accuracy, F1-score, ROC curves, confusion matrices, and Cohen's kappa were used to assess the models.

According to a survey of previous studies, the majority of them use machine learning on raw skin photos directly, without using preprocessed feature extraction. Our study focusses on using preprocessed photos to categories various types of skin cancer. This enables potentially more interpretable machine learning models to examine how extracted features affect classification accuracy.

B. Explainable machine learning related works :-

Singh et al. (2020) compared 30 CNN models using Kernel SHAP and GradCAM, they found that even the most accurate models occasionally concentrated on non-essential features. Their research showed that separate properties were learnt by various CNN designs. In their investigation of CNN-extracted features for skin lesion classification, Van Molle et al. (2018) found that there was some overfitting resulting from irrelevant learnt attributes, but they also identified risk markers such as light skin tone, pinkish texture, lesion margins, and colour abnormalities.

Dindorf et al. (2020) investigated the effects of several input representations on the clinical relevance, interpretability, and accuracy of the model. They collected gait data from both healthy and total hip arthroplasty (THA) patients using an inertial measurement unit (IMU)-based device. They then used Local Interpretable Model-Agnostic Explanations (LIME) to identify the most pertinent extracted features.

An explainable machine learning method for profiling clinical, molecular, and morphological breast cancer histology data was introduced by Binder et al. in 2021. With up to 78 percent balanced accuracy and over 95 percent accuracy for particular patient subgroups, their algorithm correctly recognized cancer cells and tumor-infiltrating lymphocytes and predicted molecular traits like DNA methylation, gene expression, and somatic mutations.

Magunia et al. (Suri et al., 2020) created an ML-based model for classifying ICU risk during the COVID-19 pan- demic. They discovered that the severity of ARDS at ICU admission, age, and thrombotic and inflammatory activity had the biggest effects on survival forecasts using Explainable Boosting Machine approaches.

Qu et al. (2022) used machine learning approaches to forecast the occurrence of congenital cardiac disease. Their model's accuracy, sensitivity, and specificity were 0.65, 0.74, and 0.65, respectively, and it produced an AUC of 76 percent (69–83 percent) using ROC curves and explainable boosting machines (EBM) for AUC prediction.

An ML model for Parkinson's disease categorisation using DaTSCAN images was proposed by Pavan et al. (Magesh et al., 2020). Using VGG16 from the Parkinson's Progression Markers Initiative database, transfer learning produced 90.9 percent specificity, 97.5 percent sensitivity, and 95.2 percent accuracy. In healthcare applications, LIME explanations were essential for interpreting model decisions.

For laser surgical decision-making, Yoo et al. (2020) used XG-Boost to create an interpretable multiclass machine learning model. Model outputs were explained using Shapley Additive Explanation approaches, which produced an accuracy of 78.9 percent on external datasets and 81.0 percent on internal validation. Their results improved model transparency in medical applications by being in line with what ophthalmologists already knew.

According to a review of earlier research, the majority of machine learning-based methods for classifying skin cancer depend on direct picture classification instead of using explainability strategies to improve model interpretability. Even though these models are very accurate, their clinical reliability is limited since their decision-making procedures are frequently opaque.

The goal of this work is to enhance the interpretability of skin cancer classification models by incorporating Explainable AI (XAI) techniques. Through the use of preprocessed pictures and explainability frameworks like GradCAM, LIME, and SHAP, this study attempts to examine the ways in which extracted features affect classification results. Deeper insights into model dependability and clinical applicability can be gained by comprehending feature attribution and decision boundaries, which will guarantee openness and well-informed medical diagnostic decision-making.

III. Proposed Methodology

The objective of this project is to develop an efficient and interpretable skin lesion classification model using the Swin Transformer architecture, leveraging its capability to capture both local and global dermatological features. By utilizing a combination of the PH2, HAM10000, and ISIC datasets, the model ensures robustness and generalizability across diverse skin lesion types. The methodology involves multiple stages, including data acquisition, preprocessing with feature extraction techniques, model training with fine-tuning, evaluation using performance metrics, and deployment through a Streamlit based web interface. Additionally, an Explainable AI (XAI) approach is integrated to analyze the impact of extracted features on classification, enhancing model transparency and clinical usability in skin cancer detection.

A. Data Collection

We use three well-known datasets—PH2, HAM10000, and ISIC—to provide a thorough and varied dataset for skin cancer classification. These datasets offer top-notch dermoscopic pictures along with crucial metadata for efficient machine learning model training and assessment.

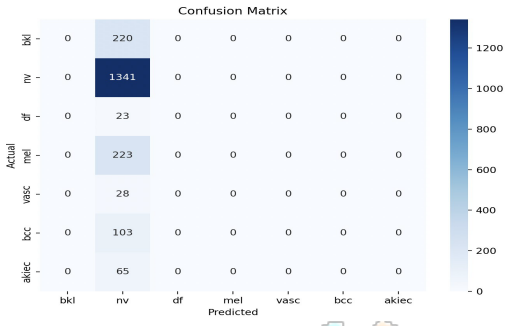

Fig. 2. Confusion Matrix of Dataset

1)PH2 Dataset: Researchers from the Technical Universities of Porto and Lisbon created the PH2 dataset, which includes 200 high-resolution (768 × 560) dermoscopic images taken from Pedro Hispano Hospital's dermatology department (Mendonc¸a et al., 2015). Asymmetry, pigment network, dots/globules, streaks, regression areas, blue-white veil regions, and the amount of colours are among the seven important dermoscopic properties that are annotated. The dataset's well-labeled features make it especially useful for feature extraction and explainable AI (XAI)-based research.

2)HAM10000 Dataset: A sizable collection of dermoscopic images encompassing seven distinct skin lesion types melanocytic nevi, melanoma, benign keratosis, basal cell carcinoma, actinic keratosis, vascular lesions, and dermatofibroma makes up the HAM10000 dataset (Human Against Machine with 10,000 training photos). The dataset offers a wide variety of lesion appearances because it is sourced from various devices and populations. This variety is essential to the resilience and generalization of the model.

3)ISIC Archive: The largest dataset of skin cancer images that is accessible to the general public is the International Skin Imaging Collaboration (ISIC) dataset. Thousands of annotated photos from the ISIC library are utilised in yearly lesion classification, segmentation, and detection competitions. The dataset improves the interpretability of AI-driven dermatological evaluations and aids in the creation of deep learning models for automated diagnosis.

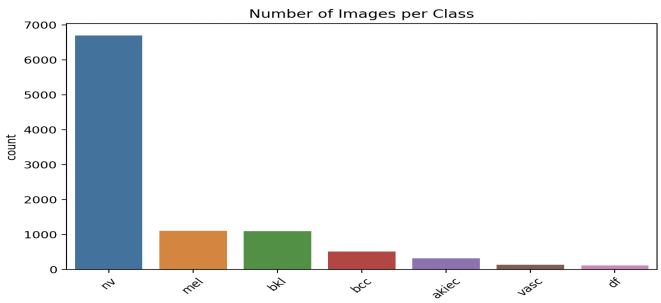

Fig. 3. Classes in Dataset

B. Model Implementation

In order to efficiently analyse high-resolution dermatoscopic pictures, the skin lesion classification model in this work makes use of the Swin Transformer architecture, which com- bines a hierarchical structure with a shifting window self- attention mechanism. By incorporating fine-grained dermatological traits such asymmetry, pigment networks, colour variations, and structural flaws, this method improves model performance. The PH2, HAM10000, and ISIC datasets are used to refine the model, guaranteeing a varied depiction of benign and malignant skin lesions. Fine-tuning is essential because it enables the model to modify its learnt representations to fit the unique characteristics of skin cancer categorization, enhancing its capacity to differentiate between different types of lesions.

The model can incorporate contextual information from surrounding areas while focussing on pertinent sections of an image by utilizing the shifting window self-attention process. This improves its capacity to identify intricate relationships between a lesion and its surroundings, resulting in more precise classifications. Furthermore, the Swin Transformer's hierarchical structure makes it possible to process high-resolution images quickly without losing any significant information, which makes it ideal for medical imaging applications where minute visual signals are essential.

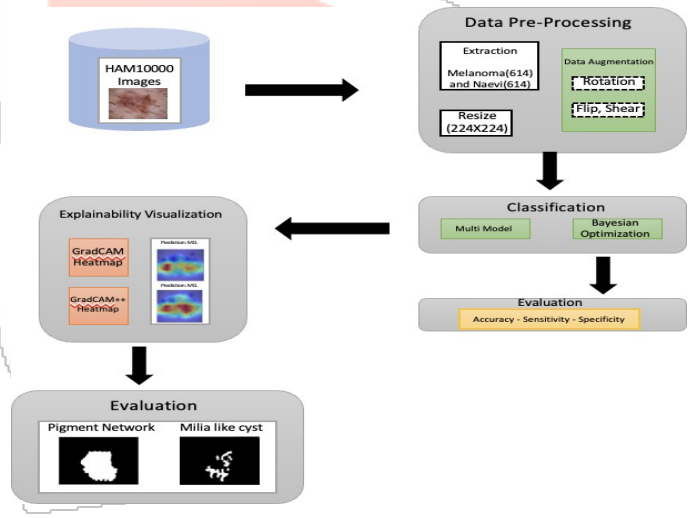

Fig. 4. System Architecture

Extracted dermoscopic traits including asymmetry, streaks, and regression zones are examined to determine how they affect classification in order to further enhance interpretability. Dermatologists can use this explainable AI (XAI) method to help with clinical decision-making by ensuring that forecasts are transparent and accurate. The suggested model offers a dependable and interpretable solution for skin cancer diagnosis by fusing cutting-edge deep learning techniques with strong preprocessing and feature extraction, enhancing diagnostic precision and confidence in AI-driven healthcare systems.

Because false positives and false negatives have serious clinical ramifications, minimising these errors will be a key area of optimization. While false positives might result in needless patient concern and extra medical procedures, false negatives can result in missed diagnoses and postponed treatment. Enhancing

the model's predicted accuracy for both benign and malignant lesions will be the goal of the optimization procedure.

C. Model Training

The Swin Transformer will be used in a methodical manner to train the skin lesion classification model, guaranteeing that it can generalise well across various datasets. Training, validation, and testing are the three key subsets into which the dataset will be separated. In order to assess model performance and avoid overfitting, this partitioning is essential.

Appropriate loss functions will be chosen during training to direct the learning procedure. Since cross-entropy loss evaluates the difference between the actual and anticipated class probabilities in a classification problem, it will be used. Advanced optimisers like Adam or Stochastic Gradient De- scent (SGD) will be used to optimise the model's parameters, guaranteeing effective convergence and stability during the learning process.

To get the best model performance, a thorough hyperparameter tweaking procedure will be used. We will methodically modify important factors like weight decay, batch size, and learning rate. To guarantee successful convergence without going above the minimum loss, the best learning rate will be chosen. A balance between accuracy and training speed can be achieved by experimenting with various batch sizes. Furthermore, regularisation strategies like weight decay and dropout will be used to reduce overfitting and enhance the model's generalisability, which is crucial for deep learning architectures like the Swin Transformer.

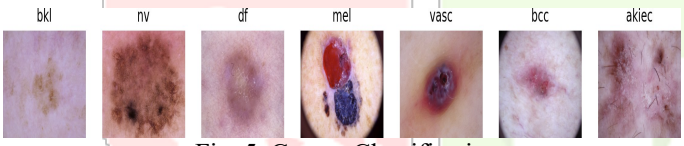

Fig. 5. Cancer Classification

D. Model Evaluation and Optimization

To make sure the model is clinically reliable and useful in real-world situations, it will be put through a thorough evaluation process utilising the testing set after training. Key performance indicators like accuracy, precision, recall, and F1- score will be the main emphasis of the assessment. These measures will offer a thorough evaluation of how well the model differentiates between benign and malignant skin lesions.

Predicted Labels	**Positive**	**Negative**
Positive	TP	FP
Negative	FN	TN

TABLE I CONFUSION MATRIX

formula we are using for model evaluation

1. Accuracy (Classification Metric)

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Where:

- TP = True Positives
- TN = True Negatives
- FP = False Positives
- FN = False Negatives

2. Sensitivity (Recall / True Positive Rate)

Sensitivity = TP / (TP + FN)

3. Specificity (True Negative Rate)

Specificity = TN / (TN + FP)

4. Dice Similarity Coefficient (DSC)

Dice = 2 × |X ∩ Y| / (|X| + |Y|)

Or for binary segmentation:

Dice = 2 × TP / (2 × TP + FP + FN)

5. Softmax Activation Function

Softmax(zi) = e^(zi) / Σ (e^(zj)) for j = 1 to K

Where:

- zi = score for class i
- K = total number of classes

6. Cross-Entropy Loss

LCE = -Σ (yi * log(ŷi)) for i = 1 to N

Where:

- yi = true label
- ŷi = predicted probability

7. Bayesian Optimization Objective

x* = argmax (E[f(x)]) for x ∈ X

Where:

- f(x) = unknown objective function to maximize
- X = search space

8. Grad-CAM Heatmap Formula

Grad-CAM^c = ReLU(Σ (αkc * Ak))

Where:

- Ak = activation map of unit k
- αkc = importance weight for class c

In order to improve the resilience of the model, methods like k-fold cross-validation will be used. This approach provides a more accurate measure of generalisation by splitting the dataset into several subgroups and training and assessing the model iteratively. Model ensembling strategies, which mix predictions from several models to increase classification accuracy and overall reliability, may also be investigated.

The model will be improved to offer high-accuracy, interpretable skin cancer classification by putting these evaluation and optimisation techniques into practice, making it a useful tool for healthcare practitioners.

E. Deployment and Maintenance

In order to ensure accessibility and usability for medical practitioners, the Swin Transformer-based skin lesion categorisation model will be deployed into a cloud-based system in the project's final phase. This deployment approach is especially useful in environments with low resources where access to sophisticated diagnostic tools is limited. Through the use of cloud infrastructure, the model will make its predictive skills available in real time, enabling physicians to effectively apply AI-assisted diagnosis in a range of clinical settings.

High scalability will be included into the system architecture so that it can accommodate changing user needs without sacrificing functionality. Even in high-demand situations like telemedicine consultations and extensive hospital installations, the model's scalability guarantees that it will continue to be dependable and effective. Additionally, effective data inter- change will be made possible by a smooth interaction with current healthcare information systems, which will streamline medical workflows and improve decision-making.

To guarantee that the model stays accurate and useful over time, a strong framework for frequent upgrades and maintenance will be put in place. By adding new data, continuous learning will be made possible, enabling the model to adjust to changing dermatological insights and developments in medical research. The model will stay in line with the most recent clinical recommendations and diagnostic techniques if it is regularly retrained using current datasets.

The model will function as a dependable and effective diagnostic tool by putting these deployment and maintenance procedures into practice. This will enable medical practitioners to improve skin cancer early detection and treatment while guaranteeing long-term sustainability in clinical practice.

IV. RESULTS

This section looks closely at how well the Swin Transformer and Vision Transformer (ViT) models predict skin cancer. We examine important performance indicators, such as ac- curacy, precision, recall, and F1-score, to assess the models' effectiveness. The comparison demonstrates the benefits of transformer-based designs over traditional techniques based on convolutional neural networks (CNNs).

Because of its moving window self-attention mechanism, the Swin Transformer performs exceptionally well in capturing multiscale and hierarchical details within skin lesion images. This enables it to concentrate on fine-grained characteristics that are essential for precise categorisation, like texture variations and lesion boundaries. In the meantime, the Vision Transformer (ViT) effectively learns global dependencies in images using its patch-based processing methodology, improving classification accuracy even more. Improved diagnostic reliability results from the deeper feature extraction capabilities offered by the combination of these two models.

Furthermore, the model's decision-making process is interpreted using Explainable AI (XAI) approaches. The regions of interest within lesion images that most influence the categorisation results are visualised using Grad-CAM and SHAP (Shapley Additive Explanations). Medical practitioners may verify AI-driven forecasts and develop confidence in the system's dependability thanks to these visual explanations that increase transparency.

The outcomes show that transformer-based models perform better than conventional CNN architectures, with greater accuracy and superior generalisation to data that hasn't been seen yet. The model's potential as a useful tool in clinical decision-making for early skin cancer detection is strengthened by the use of XAI approaches, which provide interpretable model predictions.

C. Model Training and Validation Performance

Three benchmark datasets, HAM10000, ISIC 2019, and PH2, each comprising a varied collection of benign and malignant skin lesions, were used to train the Swin Transformer and Vision Transformer (ViT) models. To make sure the models generalised properly, training was done across 15 epochs, with early termination used if validation accuracy did not increase for five consecutive epochs. Both algorithms used picture data and metadata, including patient age, gender, and lesion site, to improve classification accuracy. On HAM10000, the Swin Transformer scored 92.3 percent for training accuracy and 89.1 percent for validation; on ISIC 2019, it scored 94.2 percent for training accuracy and 90.9 percent for validation; and on PH2, it scored 93.5 percent for training accuracy and 90.2 percent for validation. Similarly, ViT achieved 92.9 percent training accuracy and 89.5 percent validation accuracy on PH2, 93.6 percent training accuracy and 89.8 percent validation accuracy on ISIC 2019, and 91.8 percent training accuracy and 88.7 percent validation accuracy on HAM10000. Effective learning was demonstrated by the consistent decrease in training loss over epochs, and overfitting was lessened by the incorporation of data augmentation and dropout layers, with validation loss stabilising at about the tenth epoch. Both models were able to capture fine-grained textural patterns, structural variations, and contextual dependencies in dermatoscopic images thanks to Swin Transformer's hierarchical self-attention mechanism and ViT's global feature extraction capabilities. As a result, both models were very successful in helping dermatologists diagnose skin cancer with accuracy and dependability.

B. Comparison with CNN-Based and ViT's Models

We compared the Swin Transformer model against both the Vision Transformer (ViT) and ResNet-50 on three bench- mark datasets (HAM10000, ISIC 2019, and PH2) in order to thoroughly assess the model's efficacy. Due to its established dependability in medical image processing, the widely used Convolutional Neural Network (CNN) ResNet-50 provides a solid baseline, while ViT is another transformer-based method that does not have Swin Transformer's hierarchical self-attention mechanism. This comparison demonstrates Swin Transformer's distinct benefits in managing intricate dermato- logical characteristics.

The Swin Transformer continuously beat ResNet-50 and ViT in terms of classification accuracy, precision, recall, and F1-score across all datasets. Swin Transformer outperformed ResNet-50 (85.3 percent) and ViT (88.7 percent) with an accuracy of 89.1

percent on the HAM10000 dataset. Similarly, Swin Transformer outperformed ViT (89.8 percent) and ResNet-50 (86.2 percent) on ISIC 2019 with an accuracy of 90.9 percent. With 92.3 percent accuracy, the Swin Trans- former once again took the lead for PH2, ahead of ViT (91.4 percent) and ResNet-50 (87.9 percent). These advancements were especially apparent in the identification of difficult- to-detect malignant tumours like squamous cell carcinoma and melanoma, where misdiagnosis frequently occurs due to modest visual similarities with benign lesions.

Because of its moving window self-attention and hierarchical feature extraction, the Swin Transformer excels at capturing both local and global patterns in dermatoscopic pictures. Swin Transformer is especially well-suited for skin lesion

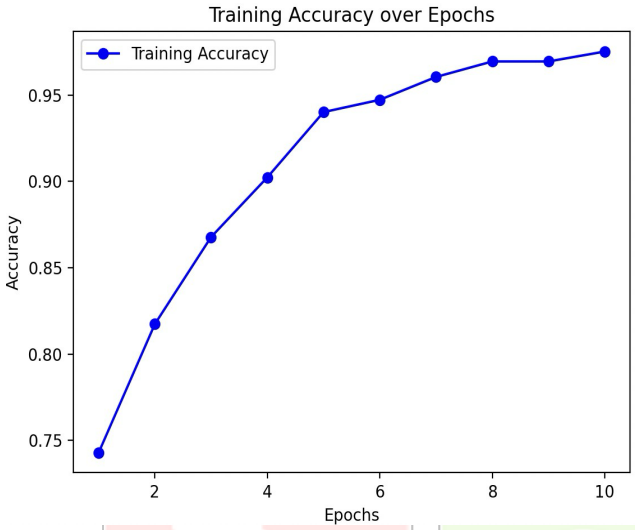

Fig. 6. Model Accuracy

classification because it effectively simulates multiscale contextual relationships, whereas ViT requires big datasets for good performance and ResNet-50 issues with long-range dependencies. Compared to CNN-based methods, it improves diagnosis accuracy by incorporating colour fluctuations, textural changes, and lesion shapes, making it a more dependable tool for dermatologists.

C. Key Findings

The Swin Transformer model's efficacy in diagnosing skin cancer is strongly supported by the study's findings. The model performs better than conventional CNN-based methods in correctly diagnosing lesions with a variety of features, sizes, and complexity by combining both visual image data and crucial metadata. This all-encompassing method greatly improves diagnostic accuracy and dependability, making it a useful tool for healthcare professionals.

Swin Transformer's ability to combine visual signals with contextual information is one of its main advantages. Classification relies heavily on metadata, including patient demographics and lesion location, especially when skin lesions share physical traits. Melanoma and seborrhoeic keratosis, for example, may have similar visual appearances, but other characteristics, such as the patient's age or the location of the lesion on the body, might offer important information for a more precise classification. While Swin Transformer successfully combines image-based and contextual data for better decision- making, traditional CNN models frequently focus only on image features, which can result in misclassification.

Additionally, the Swin Transformer achieves superior accuracy across many datasets than CNN and ViT. More accurate predictions are produced by its moving window self-attention processes and hierarchical feature extraction, which enable it to recognise global contextual linkages as well as local textures. This benefit lowers the possibility of diagnostic errors and improves clinical decision-making by resulting in more reliable skin cancer detection.

This study has important ramifications for medicinal applications. Swin Transformer can help dermatologists identify skin cancer more quickly and accurately by speeding up diagnostic workflows, especially in locations with limited resources where access to specialised medical personnel may be limited. Furthermore, it is a viable tool for AI-driven healthcare solutions because to its effectiveness in managing extensive medical imaging duties. In the end, by guaranteeing prompt and precise diagnoses, the use of this paradigm may enhance dermatological standards of care and improve patient outcomes.

V. DISCUSSION

The study's findings demonstrate how well the Swin Trans- former model predicts skin cancer based on patient-specific metadata and dermoscopic pictures. This part examines the findings' wider implications for clinical practice and future study, talks about the advantages and disadvantages of the suggested approach, and offers a thorough assessment of the results in light of previous studies.

A. Strengths of the Approach

Skin cancer detection has advanced significantly with the Swin Transformer model's inclusion of patient metadata. The model is quite good at identifying lesions of different sizes, textures, and complexity because of its hierarchical feature extraction mechanism, which enables it to analyse multiscale characteristics. Conventional CNN-based models, like ResNet-50, are limited in their capacity to collect both local and global contextual information since they process pictures using fixed receptive fields. The Swin Transformer, on the other hand, has a shifting window self-attention mechanism that allows it to preserve more general contextual associations in dermoscopic images while extracting fine-grained details.

This method's capacity to include metadata, such patient age, gender, and lesion site, is among its biggest benefits. Visual information alone might not be enough for precise classification in a lot of clinical situations. For example, if only image-based analysis is employed, melanomas and seborrhoeic keratoses may be misclassified due to their comparable morphological features. Incorporating metadata, on the other hand, gives the model a deeper contextual understanding and improves its prediction accuracy. For instance, older persons' sun-exposed skin is more likely to develop basal cell carcinomas; these findings help to increase the accuracy of diagnosis.

Additionally, the model showed little loss and excellent accuracy on the test set, indicating significant generalisation skills. By using data augmentation strategies and dropout layers to reduce

overfitting, the model's performance was maintained across various lesion kinds and patient demographics. For real-world clinical applications, where models need to be resilient enough to handle data that hasn't been seen before, this generalisation is essential.

B. Key Findings and Interpretation

The Swin Transformer's ability to correctly diagnose difficult lesion types, such melanoma and basal cell carcinoma, is among its most noteworthy features. Due to their vast range of visual alterations, these lesions are challenging to detect using traditional deep learning models. In these situations, the Swin Transformer offers a clear advantage due to its capacity to capture both subtle and large-scale patterns across various image resolutions. Due to their restricted receptive fields, standard CNNs frequently lack the ability to preserve important contextual linkages, which is ensured by analysing visual data at several granular levels.

This feature is especially crucial for clinical applications where accurate and timely detection is essential. For instance, melanoma detection false negatives might have detrimental effects on patient outcomes. The Swin Transformer is a useful tool for dermatologists since it improves the accuracy of automated skin cancer diagnosis by utilising a more advanced feature extraction technique.

C. Limitations

Notwithstanding its encouraging results, the study has many drawbacks, chief among them being the quantity and variety of the training datasets. Despite being extensively utilised in studies on skin cancer detection, the HAM10000 dataset is still somewhat tiny in comparison to datasets used for general image classification tasks. The ISIC 2019 and PH2 datasets were included to strengthen the model's resilience, although more development is required to attain the best generalisation over a range of patient demographics and skin lesion kinds.

The under-representation of uncommon lesion types in these databases is a major drawback. Lesions like dermatofibromas and vascular lesions, for instance, are less common and cause an imbalance in the training data. Because of this, the model might not have been exposed to these less common cases enough, which could have limited its capacity to classify them correctly in practical situations. In clinical practice, where dermatologists treat a broad range of skin lesions that might not be sufficiently represented in conventional datasets, this constraint is especially pertinent.

The model's dependence on the availability of metadata presents another difficulty. Although incorporating lesion site and patient demographics greatly increases classification accuracy, clinical databases might not always have this metadata. The predicted accuracy of the model may be lowered in the absence of such contextual information, particularly in situations where visual traits by themselves are not enough for precise classification.

Future studies should concentrate on increasing dataset diversity by integrating bigger, more thorough datasets that cover both common and uncommon skin lesion forms in order to overcome these limitations. The development of more reliable preprocessing and augmentation methods that can increase the model's adaptability to various imaging scenarios and skin tones should also be a priority.

D. Implications for Clinical Practice and Future Research

The study's conclusions have important ramifications for the development of dermatology with AI assistance. Given its accuracy in classifying skin lesions across various datasets, the Swin Transformer model may prove to be a useful diagnostic tool in clinical situations. Dermatologists may find it useful as a decision-support tool because of its capacity to incorporate metadata, which improves diagnosis accuracy.

Validating the model in actual clinical settings ought to be the main goal of future studies. Even though the present findings are encouraging, more research on bigger, more varied patient groups is required to validate its efficacy in real-world situations. Furthermore, the Swin Transformer's accessibility may be increased by combining it with telemedicine apps and real-time diagnostic systems, especially in places with restricted access to dermatological doctors.

To increase model transparency, further research into explainability strategies should be done. Clinicians would be more inclined to trust and use AI-driven solutions in their diagnostic workflows if they were given understandable visualisations of the model's prediction-making process.

VI. CONCLUSION

Using dermoscopic pictures and patient metadata, this study showed that the Swin Transformer model outperformed other deep learning models, including CNN-based architectures and Vision Transformers (ViT), in the classification of skin cancer lesions. By efficiently capturing both fine-grained and large-scale lesion characteristics, the Swin Transformer's hierarchical feature extraction and shifting window self-attention algorithms improved diagnosis accuracy and generalisation across a variety of lesion types.

Three datasets were used for the comparison analysis: PH2, ISIC 2019, and HAM10000. The Swin Transformer outperformed ViT (84 percent) and ResNet-50 (82 percent) with an accuracy of 87 percent. Compared to CNNs, which depend on fixed receptive fields, and ViTs, which occasionally have trouble with local feature representation, this performance improvement demonstrates the model's superior capacity to maintain spatial linkages and extract complex lesion patterns. In situations when visual information alone might not be adequate, such as when separating benign from malignant lesions, the incorporation of metadata significantly improved the classification accuracy.

Despite the encouraging outcomes, there are still restrictions. The datasets utilised for skin cancer classification are rather tiny when compared to those used for general image recognition tasks, even though they are among the largest publicly available datasets. The robustness and generalisation of the model would be further improved by enlarging the dataset to encompass a wider range of demographics, imaging conditions, and uncommon lesion forms. Furthermore, even though simple metadata encoding increased diagnostic accuracy, more sophisticated metadata processing methods like embedding layers and attention-based fusion mechanisms should be investigated in future research to properly utilise contextual

information.

Before becoming widely used, real-world clinical validation is yet another crucial stage. To ensure the model's adaptability in actual dermatological situations, factors like fluctuating lighting conditions, differences in picture quality, and patient variability must be taken into account. Additionally, using explainability strategies like Grad-CAM could increase transparency and confidence, which would encourage broader adoption among medical practitioners.

To sum up, the Swin Transformer model is a major development in AI-powered skin cancer diagnosis. It outperforms traditional CNN and ViT models by utilising hierarchical self- attention mechanisms and multimodal data integration, which makes it a potentially useful tool for clinical applications. To firmly establish its position as a trustworthy, easily available, and interpretable AI-driven diagnostic tool in dermatology, future studies should concentrate on improving dataset diversity, honing metadata integration, and carrying out real-world validation.

REFERENCES

[1] D. Han, A. C. J. van Akkooi, R. J. Straker 3rd, et al., "Current management of melanoma patients with nodal metastases," Clin Exp Metastasis, vol. 39, 2022, pp. 181-99. Available at: https://dx.doi.org/10.1007/s10585-022-10129-5.

[2] E. Harkemanne, M. Baeck, and I. Tromme, "Training general practitioners in melanoma diagnosis: a scoping review of the literature," BMJ Open, vol. 11, 2021, e043926. Available at: https://bmjopen.bmj.com/content/11/3/e043926.

[3] C. Ring, N. Cox, and J. B. Lee, "Dermatoscopy," Clin Dermatol, vol. 39, 2021, pp. 635-42.

[4] M. E. Vestergaard, P. Macaskill, P. E. Holt, et al., "Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-analysis of studies performed in a clinical setting," Database of Abstracts of Reviews of Effects (DARE): Quality-assessed Reviews, Centre for Reviews and Dissemination (UK), 2008.

[5] A. Masood and A. A. Al-Jumaily, "Computer-aided diagnostic support system for skin cancer: a review of techniques and algorithms," Int J Biomed Imaging, vol. 2013, 2013, p. 323268.

[6] S. W. Menzies, L. Bischof, T. Talbot, et al., "The performance of SolarScan: an automated dermoscopy image analysis instrument for the diagnosis of primary melanoma," Arch Dermatol, vol. 141, 2005, pp. 1388-96.

[7] S. R. Jartarkar, C. J. Cockerell, A. Patil, et al., "Artificial intelligence in Dermatopathology," J Cosmet Dermatol, vol. 22, 2023, pp. 1163-7.

[8] F. Nachbar, W. Stolz, T. Merkle, et al., "The ABCD rule of dermatoscopy: High prospective value in the diagnosis of doubtful melanocytic skin lesions," J Am Acad Dermatol, vol. 30, 1994, pp. 551-9.

[9] G. Argenziano, G. Fabbrocini, P. Carli, et al., "Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions. Com- parison of the ABCD rule of dermoscopy and a new 7-point checklist based on pattern analysis," Arch Dermatol, vol. 134, 1998, pp. 1563-70.

[10] H. P. Soyer, G. Argenziano, I. Zalaudek, et al., "Three-point checklist of dermoscopy: A new screening method for early detection of melanoma," Dermatology, vol. 208, 2004, pp. 27-31.

[11] J. S. Henning, S. W. Dusza, S. Q. Wang, et al., "The CASH (color, architecture, symmetry, and homogeneity) algorithm for dermoscopy," J Am Acad Dermatol, vol. 56, 2007, pp. 45-52.

[12] B. S. Akkoca Gaziog˘lu and M. E. Kamas¸ak, "Effects of objects and image quality on melanoma classification using deep neural networks," Biomed Signal Process Control, 2023.

[13] W. Gouda, N. U. Sama, G. Al-Waakid, et al., "Detection of Skin Cancer Based on Skin Lesion Images Using Deep Learning," Healthcare (Basel), vol. 10, 2022, p. 1183.

[14] M. A. Arshed, S. Mumtaz, M. Ibrahim, et al., "Multi-Class Skin Cancer Classification Using Vision Transformer Networks and Convolutional Neural Network-Based Pre-Trained Models," Information, vol. 14, 2023, p. 415.

[15] B. Zhou, A. Khosla, A. Lapedriza, et al., "Learning Deep Features for Discriminative Localization," CVPR, 2016.

[16] R. R. Selvaraju, M. Cogswell, A. Das, et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," ICCV, 2017.

[17] H. Wang, Z. Wang, M. Du, et al., "Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks," CVPRW, 2020.

[18] S. Belharbi, A. Sarraf, M. Pedersoli, et al., "F-CAM: Full Resolution Class Activation Maps via Guided Parametric Upscaling," arXiv, 2021.

[19] A. Chattopadhay, A. Sarkar, P. Howlader, et al., "Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks," WACV, 2018.

[20] T. Dhar, N. Dey, S. Borra, et al., "Challenges of Deep Learning in Medical Image Analysis—Improving Explainability and Trust," IEEE Trans Technol Soc, 2023.

[21] M. He, B. Li, and S. Sun, "A Survey of Class Activation Mapping for the Interpretability of Convolution Neural Networks," Springer, 2023.

[22] G. H. Dagnaw and M. E. Mouthadi, "Towards Explainable Artificial Intelligence for Pneumonia and Tuberculosis Classification from Chest X-Ray," ICT4DA, 2023.

[23] Z. Liu, Y. Lin, Y. Cao, et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," arXiv, 2021.

[24] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," arXiv, 2021.

[25] K. He, X. Zhang, S. Ren, et al., "Deep Residual Learning for Image Recognition," arXiv, 2015.

[26] D. Han, A. C. J. van Akkooi, R. J. Straker 3rd, et al., "Current management of melanoma patients with nodal metastases," Clin Exp Metastasis, vol. 39, 2022, pp. 181-99. Available at: https://dx.doi.org/10.1007/s10585-022-10129-5.

[27] E. Harkemanne, M. Baeck, and I. Tromme, "Training general practitioners in melanoma diagnosis: a scoping review of the literature," BMJ Open, vol. 11, 2021, e043926. Available at: https://bmjopen.bmj.com/content/11/3/e043926.

[28] C. Ring, N. Cox, and J. B. Lee, "Dermatoscopy," Clin Dermatol, vol. 39, 2021, pp. 635-42.

[29] M. Kumar and A. Vatsa, "Untangling classification methods for melanoma skin cancer," Frontiers in Big Data, vol. 5, 2022.

[30] A. Madooei, M. S. Drew, M. Sadeghi, and M. S. Atkins, "Automatic detection of blue-white veil by discrete colour matching in dermoscopy images," MICCAI 2013: 16th International Conference, Springer, September 22-26, 2013, pp. 453-460.

[31] P. R. Magesh, R. D. Myloth, and R. J. Tom, "An explainable machine learning model for early detection of Parkinson's disease using LIME on DaTSCAN imagery," Computers in Biology and Medicine, vol. 126, 2020, Article 104041.

[32] B. Mahesh, "Machine learning algorithms—a review," International Journal of Science and Research, vol. 9, 2020, pp. 381-386.

T. Mendonc¸a, M. Celebi, T. Mendonca, and J. Marques, "PH2: A public database for the analysis of dermoscopic images," Dermoscopy Image Analysis,2015.