Innovative Melanoma Diagnosis: Harnessing VI Transformer Architecture

Sreelakshmi Jayasankar[1]*, T. Brindha[2]
Research Scholar, Department of Computer Science and Engineering[1]
Associate Professor, Department of Information Technology[2]
Noorul Islam Centre for Higher Education, Tamil Nadu, India[1, 2]

***Abstract*—Melanoma, the most severe type of skin cancer, ranks ninth among the most prevalent cancer types. Prolonged exposure to ultraviolet radiation triggers mutations in melanocytes, the pigment -producing cells responsible for melanin production. This excessive melanin secretion leads to the formation of dark-colored moles, which can evolve into cancerous tumors over time and metastasize rapidly. This research introduces a Vision Transformer, revolutionizes computer vision architecture by diverging from traditional convolutional neural networks, employing transformer models to handle images as sequences of flattened, spatially-structured patches. The dermoscopy images sourced from the Kaggle repository, an extensive online database known for its diverse collection of high-quality medical imagery is utilized in this study. This novel deep learning model for melanoma classification, aiming to enhance diagnostic accuracy and reduce reliance on expert interpretation. The model achieves an accuracy of 96.23%, indicating strong overall correctness in classifying both Benign and Malignant cases. Comparative simulation of the proposed method against other methods in skin cancer diagnosis reveal that the suggested approach attains superior accuracy. These findings underscore the efficacy of the system in advancing the field of skin cancer diagnosis, offering promising prospects for enhanced accuracy and efficacy in clinical settings.**

Keywords*—Vision transformer; melanoma; convolutional neural networks; deep learning model; transformer encoder; dermoscopy image

I. INTRODUCTION

Prevalence of skin-related diseases has surged in recent years surpassing common conditions like hypertension and obesity [1]. Skin disorders account for approximately 12.4% of cases, affecting roughly one in every three individuals [2]. A concerning trend, a yearly increase of 1-2% in recorded skin diseases. Among these, melanoma stands out as the most aggressive form of skin cancer, capable of metastasizing through the lymphatic system and bloodstream to distant parts of the body. Melanoma arises from melanocytes, pigment-producing cells situated at the junction of the epidermis and dermis [3]. These cells are responsible for melanin production, when melanocytes undergo abnormal mutation, melanoma develops. This malignant condition poses a significant health risk due to its potential for rapid spread and invasive behaviour, underscoring the importance of early detection and effective treatment strategies.

Melanoma, a relatively uncommon form of skin cancer, poses a significant threat to mortality rates [4]. Although imaging studies can detect metastatic spread, the disease frequently goes undiagnosed until it progresses to an advanced stage or spreads to the bloodstream or lymph nodes [5]. It is essential to develop efficient computational techniques for early melanoma diagnosis. The five main forms of melanoma are nodular, lentigo maligna, Acral lentiginous, Subungual, and superficial spreading [6]. Each has a unique set of symptoms, interestingly amelanotic melanoma is a distinct subtype that occurs in people of different skin tones.

Conventional methods of melanoma diagnosis have limitations in terms of accuracy, accessibility, and scalability. Moreover, the increasing prevalence of melanoma underscores the urgent need for efficient and reliable diagnostic tools to address this public health challenge. In recent years, artificial intelligence (AI) and machine learning (ML) techniques have spurred the development of automated melanoma detection systems. These systems leverage computer vision algorithms, deep learning (DL) architectures, and large-scale datasets to analyze dermoscopy images and distinguish between benign and malignant lesions [7].

This paper aims to explore the current landscape of melanoma detection methodologies, highlighting the challenges and opportunities in this evolving field. Additionally, we present a comprehensive review of recent advancements in AI-based melanoma detection techniques, focusing on their strengths, limitations, and potential for clinical integration. The contributions of this work can be outlined as follows:

- Development of classification model based on DL aimed at effectively detecting and classifying melanoma, with a focus on enhancing detection performance.
- Assessment of the algorithm's performance using established benchmark metrics for evaluation.
- To assess its efficacy and explore its methodological strengths, compare the proposed model with existing models.

The rest of the paper is organized as follows: In Section II, a summary of literature is provided, highlighting areas that indicate a need for more investigation. In Section III, the methodology is explained in depth. Section IV goes into great detail about the results that the suggested strategy produced. A

discussion is provided in Section V and finally, a summary of the findings is included in Section VI, which gives a conclusion to the paper.

II. Literature Review

Melanoma detection has emerged as a significant global concern, drawing the interest of researchers worldwide who seek optimal methods for early identification of skin abnormalities to mitigate their progression. Numerous research endeavors have been initiated and continue to evolve in this field, aiming to improve patient outcomes and enhance the efficacy of medical interventions. This section provides a comprehensive overview of various research initiatives focused on melanoma detection.

Adla et al. [8] proposed a DL model for skin lesion detection. Tsallis entropy was utilized to identify the affected lesion areas in the dermoscopy images. Capsule Network in conjunction with class attention layer and Adagrad optimizer was utilized to extract features from the segmented lesions. The Convolutional Sparse Autoencoder, which was based on the Swallow Swarm Optimization algorithm, did the classification. The detection method exhibited limitations, particularly in its performance when presented with noisy images.

A CNN-based framework was presented by Shorfuzzaman et al. [9] to identify melanoma skin cancer. The final predictions were produced by a meta-learner, which incorporated all of the predictions from the sub models. The evaluation results demonstrated 95.76% accuracy in the ensemble model. A notable limitation of the paper lay in the extended duration necessitated for training, indicating a potential challenge in terms of resource allocation and efficiency within the framework.

To categorize the image samples of skin lesions, a framework was proposed by Khan et al. [10], consisting of two modules—the categorization and the localization of skin lesions. Transfer learning was used in the classification module to retrain a pre-trained DenseNet201 model on the segmented lesion images. The distribution stochastic neighbor embedding technique was used to downsample the features that were retrieved from the two fully connected layers. Using a fused vector, the highest accuracy on the ISBI2017 was 95.26%. One significant limitation of the work was that the model's training on localized regions entailed longer time in comparison to training on raw dermoscopy images, potentially impeding the scalability and practicality of the proposed approach.

Jiang et al. [11] introduced DRANet, a lightweight deep learning framework, for the classification of 11 types of skin diseases using real histopathological images. DRANet was the incorporation of a Squeeze Excitation Attention, which directed the framework's focus towards key areas crucial for identifying specific skin diseases. By employing stacked modules, the framework enhanced its capacity to learn from high-level features. Despite achieving an accuracy of 86.8%, the proposed approach was constrained by its inability to effectively diagnose images of poor quality.

Yacin Sikkandar et al. [12] introduced a model for skin lesion diagnosis, with an Adaptive NeuroFuzzy classifier merging a GrabCut algorithm. The model underwent simulation utilizing a benchmark ISIC dataset. Two significant limitations of the paper were the prolonged training time and the demand for substantial computational resources. The method relied on a large volume of data, posing a challenge in terms of data acquisition and processing.

Zghal et al. [13] sought to devise a straightforward model for detecting skin lesions from dermoscopy images, leveraging ABCD rules. Their approach consisted of five sequential stages: acquisition, pre-processing involving noise elimination and contrast enhancement techniques, and ultimately, classification via Total Dermoscopy Value computation. However, a notable constraint of their algorithm was its reliance on a substantial dataset for learning, which may not always be accessible.

Alwakid et al. [14] proposed a DL method for extracting a lesion zone in skin cancer diagnosis. ESRGAN was utilized to enhance image quality by generating high-resolution versions of low-resolution images. Melanoma and non-cancerous lesions could be distinguished using a ViT-based architecture suggested by Cirrincione et al. [15]. Based on the ISIC dataset, the suggested predictive model was evaluated with an accuracy of 94.8%.

An automated image-based method using ML classification techniques was presented by Inthiyaz et al. [16] for the diagnosis and classification of skin diseases. Their approach used the softmax classifier algorithm to identify images by leveraging Convolutional Neural Networks (CNNs). Six prevalent skin disorders were represented by images in the dataset, which showed different facial skin ailments. Their method showed significant effectiveness in skin condition detection and diagnosis with an obtained accuracy of 87%.

A. Research Gap

The research encounters challenges in performance attributed to the intricate visual attributes inherent in skin lesion images, characterized by diverse features and ambiguous boundaries. Detection accuracy notably diminishes for lesions smaller than 6mm, presenting a formidable hurdle in melanoma identification. Early melanoma symptoms often resemble benign skin conditions such as age spots and moles, underscoring deficiencies in early detection approaches. Furthermore, the subtle presentation of melanoma symptoms complicates early-stage detection for individuals, exacerbating gaps in effective detection strategies. Scarce access to medical data hampers algorithm development and training, highlighting deficiencies in data availability for research purposes. Moreover, the labor-intensive process of algorithmic development, validation, and deployment poses significant challenges, intensifying gaps in algorithmic implementation. The considerable variability in melanoma cases, including differences in size, shape, and color, poses a formidable obstacle for algorithms to achieve generalization, accentuating gaps in algorithmic robustness and adaptability.

Many models, such as those relying on capsule networks or CNN-based frameworks, exhibit performance degradation when faced with noisy or low-quality images, which is a significant drawback given the intricate visual characteristics

of melanoma lesions. Furthermore, models that require extensive training times or high computational resources, such as those employing ensemble learning or transfer learning, are impractical for real-time clinical applications where efficiency and scalability are crucial. Additionally, approaches like those utilizing ABCD rules or other manual feature extraction methods are limited by their dependence on large datasets, which are often not readily available, especially in diverse clinical environments. These limitations hinder the ability to achieve accurate and robust melanoma detection, particularly in early stages where symptoms may resemble benign conditions, thereby exacerbating gaps in effective diagnostic strategies. Consequently, a model that can address these challenges by providing reliable performance across varied image qualities, minimizing training time, and reducing dependence on extensive datasets is essential for improving melanoma detection and diagnosis. These limitations highlight the need for a more robust, adaptable, and efficient model capable of overcoming these challenges to provide accurate and timely melanoma detection.

III. Materials and Methods

The proposed methodology leverages the Vision Transformer (ViT) architecture as shown in Fig. 1, a cutting-edge approach in computer vision [17]. The Vision Transformer model was chosen for this research due to its innovative approach to image processing, which marks a significant departure from conventional convolutional neural networks (CNNs). ViT processes images by segmenting them into sequences of spatially-structured patches, which undergo linear embedding and positional encoding to preserve spatial information. This ability to handle complex, high-dimensional data makes ViT particularly well-suited for dermoscopy images, where subtle variations in color, texture, and structure are critical for accurate melanoma classification. These patches are then fed into a stack of transformer encoder blocks, allowing the model to capture global contextual dependencies and hierarchical features. Following this, the output undergoes global pooling to reduce spatial dimensions before being passed through a classification head comprising fully connected layers.

The classification phase translates spatial information into class predictions using softmax activation. Key components of the ViT architecture include patch extraction, patch embedding, and transformer encoder blocks. The self-attention mechanism within transformer encoder blocks allows the model to focus on relevant features while suppressing irrelevant ones. Residual connections and layer normalization facilitate gradient flow and stabilize input to subsequent layers. The Multi-Layer Perceptron (MLP head), the final component of the model, transforms aggregated representations from transformer encoder blocks into class predictions through fully connected layers and activation functions. Regularization techniques such as dropout layers employed to prevent over fitting. Ultimately, the methodology aims to learn representations for diverse visual patterns, fostering scalability and interpretability in image classification tasks.

A. Dataset

The data is collected from the Kaggle repository [18]. The dataset comprises a balanced collection of images representing two distinct categories: benign skin moles and malignant skin moles. Each category is represented by a folder containing 1800 images, with each image standardized to a size of 224x244 pixels. The standardized image size simplifies pre-processing tasks and ensures uniformity across the dataset, enabling straightforward integration into various ML pipelines. The balanced nature of the dataset ensures an equal representation of both benign and malignant cases, facilitating unbiased model training and evaluation. The inclusion of sample images, as depicted in the Fig. 2, provides a visual representation of the dataset contents, offering insight into the appearance and characteristics of both benign and malignant skin moles.

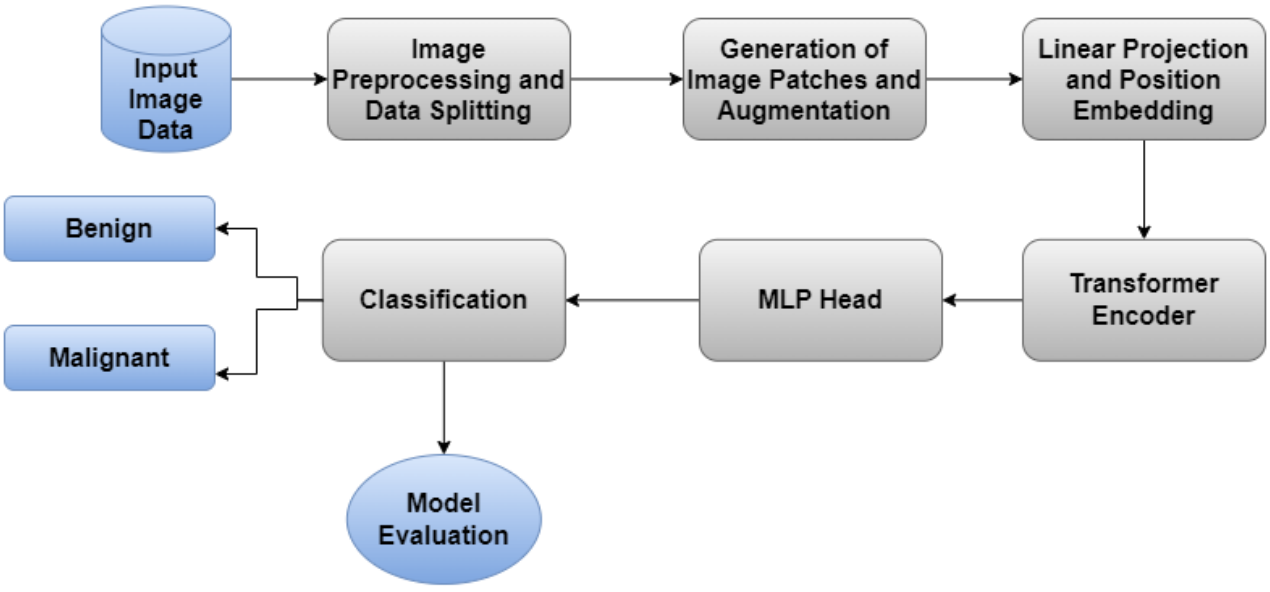

Fig. 1. Block diagram of the proposed system.

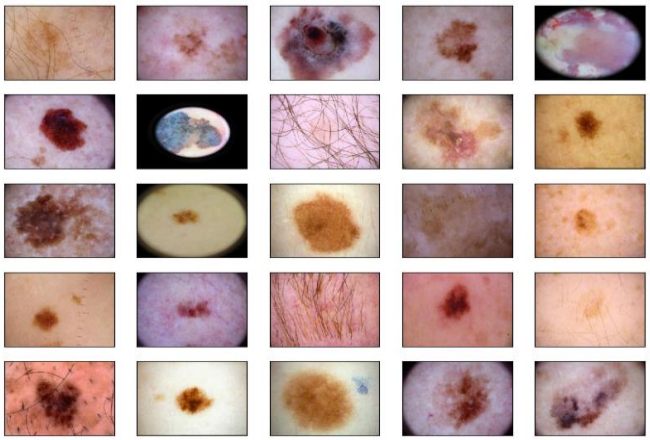

Fig. 2. Sample images from the dataset.

B. Pre-processing and Augmentation

Pre-processing involves preparing the input images for the model by standardizing various aspects to ensure consistency. One of the first steps in pre-processing is normalization, which calculates the mean and variance of the pixel values in the training images. By scaling the pixel values to a standard range, normalization minimizes the impact of intensity variations between different images. This is particularly important in medical imaging, where lighting conditions and image acquisition settings can vary significantly. Ensuring a uniform pixel intensity range helps the model focus on the relevant features of the melanoma lesions rather than being influenced by extraneous variations. Another key pre-processing step is resizing, where images are adjusted to a consistent size. This standardization ensures that all input images fit the model architecture requirements, making it easier for the model to process and analyze the images efficiently. Consistency in image dimensions also simplifies subsequent processing steps and improves the model's ability to learn from the data.

Data augmentation complements pre-processing by artificially expanding the dataset and introducing controlled variations to enhance model generalization. For melanoma detection, this often includes techniques such as horizontal flipping, rotation, and zooming. Horizontal flipping creates mirrored versions of the original images, introducing variability in image orientation. This technique helps the model learn to recognize melanoma lesions from different angles, enhancing its ability to generalize to real-world scenarios where lesions might not always be perfectly oriented. Random rotations further enrich the dataset by changing the angle at which the images are presented, making the model more adept at identifying melanoma regardless of its orientation in the image. Random zooming, on the other hand, introduces variations in the scale of the images. By simulating different distances from the lesion, zooming helps the model become proficient in detecting melanoma at various sizes and levels of detail.

C. Patch Generation

Patch generation phase extract patches from input images, a crucial process for uncovering localized features vital for a spectrum of computer vision tasks, including image segmentation, object detection, and image classification. Patch parameterization offers adaptability to various task requirements, allowing for tailored patch extraction, dictating the dimensions of the patches to be extracted. Leveraging the information including batch size, height, width, and channels alongside the specified `patch_size`, computes the number of patches extractable in both height and width dimensions, ensuring exhaustive coverage of the image data. This meticulous extraction process ensures uniform and controlled patch extraction across diverse datasets.

The extracted patches undergo reshaping process, transforming them into a 3D tensor format optimized for subsequent processing and analysis within the DL model. This transformation ensures seamless integration of patches into the larger computational framework, facilitating efficient feature extraction and model training. This serialization of patches is carried out to ensure consistency in patch generation across different model instances, facilitating seamless integration and reproducibility. Fig. 3 provides a visual comparison between a sample image and the corresponding image patches generated from it.

Fig. 3. Sample image and generated image patches.

D. Linear Projection and Positional Embedding

Patch encoding phase encode patches extracted from input images, a critical process that enriches the representation of localized features. Since transformers operate on fixed-size sequences and lack inherent understanding of spatial relationships, positional encoding is performed. During the initialization step, two important parameters are determined: the number of patches retrieved from the input images and the dimensionality of the projected feature space to which the patches are mapped. The initialization phase establishes the foundation for effective feature extraction and spatial representation within the encoded patches. Within the initialization, two sub layers are instantiated to facilitate the encoding process. It plays a role in protecting the input patches into a higher-dimensional feature space, enabling robust feature extraction.

The primary function of this layer is to provide position embeddings for every patch, thereby capturing essential spatial information within the encoded representation. By creating position indices for the patches, this technique makes sure that every patch is associated with a unique position. The input patches are then projected into the feature space that has more dimensions. The position embeddings generated are then added to the projected patches, effectively incorporating spatial

information into the encoded representation. The resulting encoded patches, enriched with both feature and positional information. The transformer processes these embeddings through multiple layers of attention mechanisms and feed forward networks.

The transformer encoder is a crucial component of the ViT architecture, responsible for processing and extracting features from input patches. It consists of several layers, each containing a set of modules such as Multi-Head Self-Attention Mechanism, Residual Connections and Layer Normalization, Feed forward Neural Network (MLP). The Fig. 4 illustrates the architectural framework of the ViT.

The mechanism of Self-attention enables the model to assign varying weights to different input patch embeddings, prioritizing relevant information while disregarding irrelevant parts. The mechanism of self-attention in the model enables differential weighting of input patch embeddings, prioritizing relevant information while downplaying irrelevant aspects. Each patch embedding undergoes linear transformation into key, query, and value vectors, which are then utilized to compute attention score. These scores are derived from the product of query and key vector, and processed with softmax function to yield attention weights. These weights are subsequently applied to the values to generate the attention output. To aid in gradient flow during training, residual connections are introduced, merging the attention output with the input and subjecting it to layer normalization. This normalization process stabilizes the output of each attention block, maintaining a consistent input range for subsequent layers. Following the attention mechanism, the output traverses through MLP consisting of two fully connected layers.

The MLP head within a ViT serves as the final stage of the model, responsible for converting the aggregated representations obtained from the transformer encoder blocks into class predictions. Typically, the MLP head comprises one or more fully connected layers, followed by an activation function. The input to this MLP head is the output derived from the last transformer encoder block, embodying the combined information extracted from the input image patches. Before being fed into the MLP head, these representations undergo flattening or global averaging. The resulting flattened or pooled representation is then passed through one or more dense layer, constituting the core of the MLP head. These layers enable the model to discern intricate non-linear relationships within the data. An activation function is applied after each fully connected layer to introduce non-linearity into the network.

E. Hardware and Software Setup

The research employed a computational setup that utilized a machine with powerful characteristics, including an Intel Core i7 processor. A powerful combination of 32GB RAM and the impressive NVIDIA GeForce GTX 1080Ti GPU. The model was implemented smoothly using the Keras library, which served as a prototype based on the Tensorflow architecture and executed using the flexible Python language. Keras, renowned for its intuitive interface and robust capabilities, played a crucial role in designing complex Neural Network structures. This framework guarantees optimal utilization of computational resources, effortlessly adapting to CPU, GPU, and TPU contexts. In order to take use of the powerful computational powers and optimise the process of training the model, the deployment was coordinated on Google Colab. Model training is made easier with this cloud-based Python notebook environment, which offers free access to powerful computing resources and supports interaction in development.

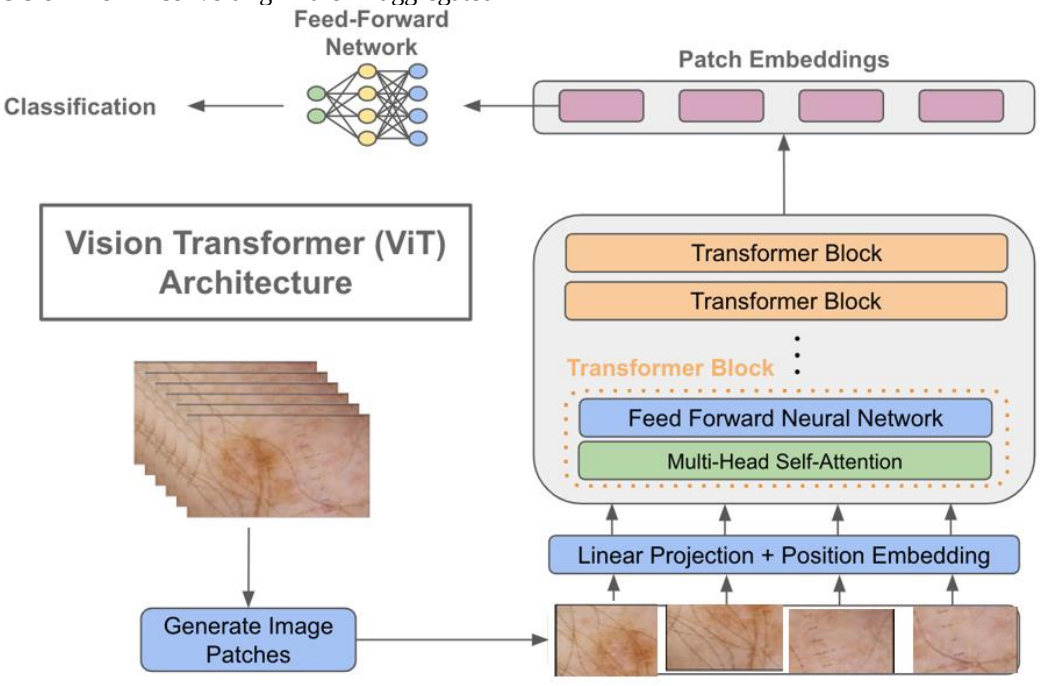

Fig. 4. Architecture of the VI transformer.

Hyper parameters are crucial configuration parameters that determine the behavior and attributes of a machine learning framework during the training phase. In contrast to the model's parameters, which are determined by the data itself, the user sets the hyper parameters prior to training. The neural network model utilizes the Adam optimizer. The training process is directed by the binary cross-entropy loss function. During the training process, the model handles input data in batches consisting of 32 samples per iteration. The training is conducted for 25 epochs, which represents the number of times the model processes the complete training dataset. The hyper parameter selections, including the optimizer, loss function, batch size, and number of epochs, determine the setup for training the neural network model. The goal is to optimise its performance in detecting melanoma. The proposed method's model configuration is presented in Table I.

TABLE I. MODEL CONFIGURATIONS

Parameters	Value
Learning rate	0.0001
Weight decay	0.00001
Image size	224
Patch size	6
Projection dimension	64
num_heads	6
transformer layers	6
mlp_head_units	[1024, 512]
Batch Size	32
Epochs	25

IV. EXPERIMENTAL RESULTS

The accuracy and loss plots are crucial for comprehending the performance and learning patterns of the proposed model. The accuracy plot visually depicts the model's ability to reliably predict data labels during training iterations on both the training and validation datasets. The alignment between the model's predictions and the actual labels is monitored to assess the model's performance throughout training.

The accuracy plot demonstrates the model's efficacy in differentiating between images containing signs of melanoma and those without, throughout the training process. Ideally, throughout the early stages, both the training and validation accuracies ought to increase simultaneously, demonstrating the model's ability to apply its knowledge beyond the training data. The trend illustrated in Fig. 5 indicates that the model is acquiring knowledge of fundamental patterns rather than only memorizing the instances presented in the training dataset.

The accuracy steadily increases from an initial value of approximately 0.8788 in the first epoch to around 0.9306 in the final epoch. This upward trajectory indicates that the model's performance improves over successive epochs as it learns from the training data. Notably, there are fluctuations in accuracy values throughout the training process, reflecting the dynamic nature of the optimization process and the model's adaptation to different patterns in the data.

A loss plot illustrates the trend of the model's loss function over different iterations or epochs during training. Fig. 6 illustrates the loss plot of the proposed model. The loss steadily decreases from an initial value of approximately 0.3062 in the first epoch to around 0.1707 in the final epoch. This downward trend signifies that the model's ability to minimize prediction errors improves over time. Lower loss values indicate better alignment between the model's predictions and the actual labels in the training data. Similar to accuracy, fluctuations in loss values are observed across epochs, reflecting the model's response to variations in the training data and optimization process.

An excellent way to assess the accuracy of the suggested model in detecting melanoma is by employing a confusion matrix. Fig. 7 presents the confusion matrix generated by the proposed model. The matrix offers a systematic summary of the model's performance by contrasting its predictions with the real labels across several classes. Essentially, it arranges the results in a tabular structure, with the rows representing the actual labels and the columns representing the predicted labels. Every individual cell in the matrix represents the number of occurrences where the model's predictions match or differ from the actual labels.

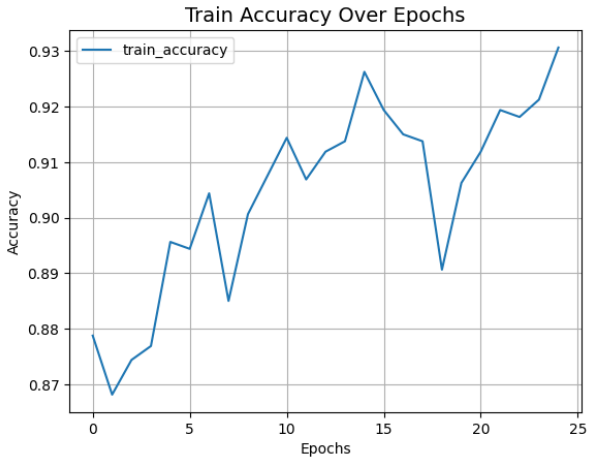

Fig. 5. Accuracy plot.

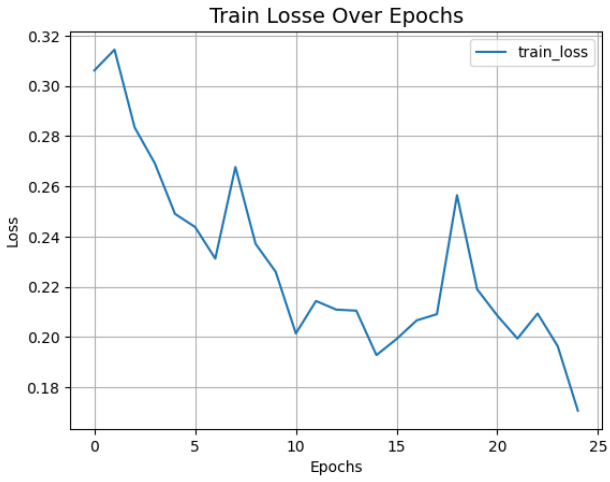

Fig. 6. Loss plot.

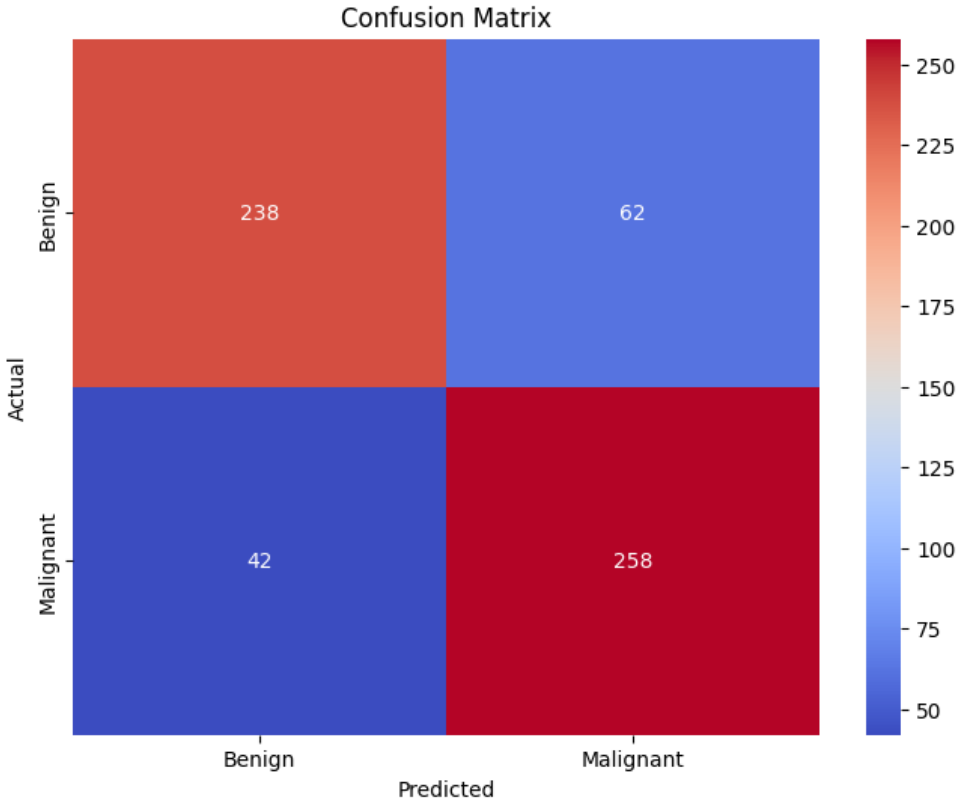

Fig. 7. Confusion matrix.

The confusion matrix is partitioned into four quadrants, where the items on the diagonal represent correct predictions and the elements off the diagonal represent cases of misclassification. This visual depiction allows for a comprehensive evaluation of the proposed model's efficacy in accurately detecting melanoma individuals. It reveals that the model accurately identifies 238 benign images as benign, but misclassifies 62 benign images as malignant. Similarly, it correctly identifies 258 malignant images as malignant, but erroneously classifies 42 malignant images as benign.

Performance metrics derived from the confusion matrix offer a thorough evaluation of the proposed model's efficacy in detecting melanoma. In order to thoroughly evaluate the efficacy and operational efficiency of the proposed model, the F1-score, accuracy, precision, and recall are the four primary metrics utilized. These measures, which are based on the concepts of False Positive (FP), False Negative (FN), True Negative (TN), and True Positive (TP), are essential for assessing the model's performance. These performance parameters have mathematical formulations that are shown in Eq. (1), (2), (3), and (4).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1-score = 2 \times \frac{precision \times Recall}{Precision+Recall} \quad (4)$$

The obtained performance metrics, as shown in Fig. 8, highlight the exceptional efficacy of the developed model. An accuracy of 96.23% indicates that the model correctly identifies melanoma cases and non-melanoma instances with a high degree of reliability. The precision of 96.63% demonstrates the model's capability to accurately predict true positive cases of melanoma, minimizing the rate of false positives. The recall, or sensitivity, at 96.98% reflects the model's effectiveness in detecting almost all actual melanoma cases, ensuring a low rate of false negatives. The F1-Score, a harmonic mean of precision and recall, is 96.80%, underscoring the model's balanced performance in terms of both identifying true cases and excluding false alarms. These metrics collectively suggest that the model is robust and highly accurate, making it a reliable tool for the detection and classification of melanoma in clinical settings.

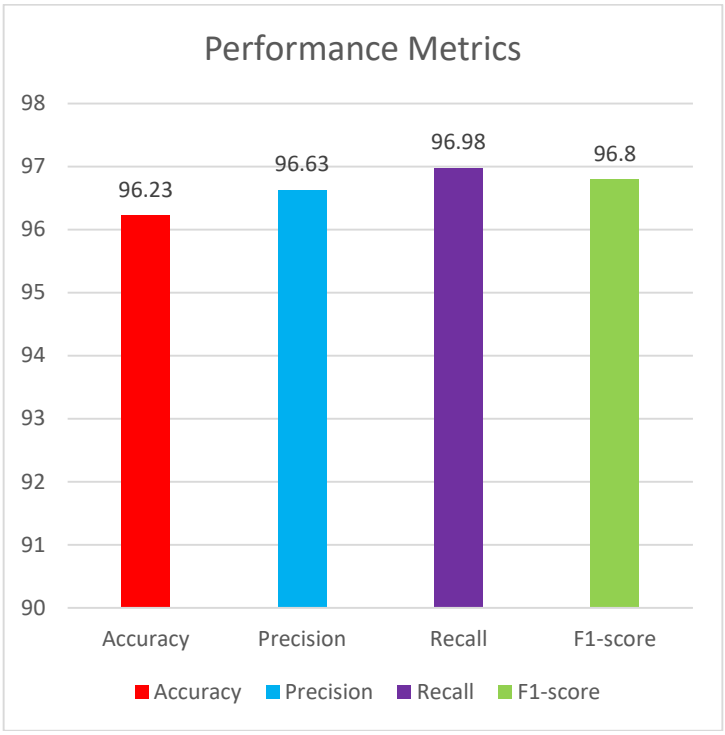

Fig. 8. Performance metrics.

Fig. 9 shows the Classification Output of the proposed system.

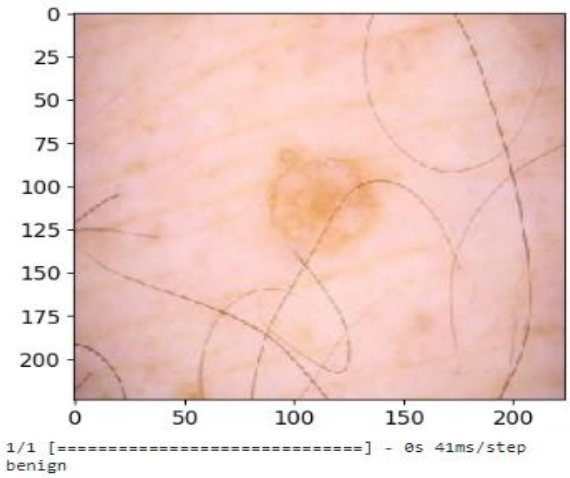

Fig. 9. Classification output of the system.

V. Discussion

The proposed Vision Transformer (ViT)-based deep learning model marks a significant advancement in the field of skin cancer diagnosis, particularly in melanoma detection. By leveraging the innovative architecture of the Vision Transformer, which processes images as sequences of spatially-structured patches rather than relying solely on convolutional layers, the model exhibits superior accuracy and robustness. As demonstrated in the Table II and Fig. 10, the ViT-based model achieves an accuracy of 96.23%, notably surpassing the performance of other state-of-the-art methodologies. Jiang et al.'s DRANet achieves an accuracy of 86.8%, while Shorfuzzaman et al.'s CNN-based stacked ensemble framework achieves a higher accuracy of 95.76%. Khan et al. employ DenseNet201 with transfer learning, achieving an accuracy of 95.26%. Inthiyaz et al.'s CNN method achieves an accuracy of 87%. This substantial improvement underscores the potential of the ViT-based model to enhance diagnostic accuracy, which is crucial for early detection and treatment of melanoma, ultimately contributing to better clinical outcomes.

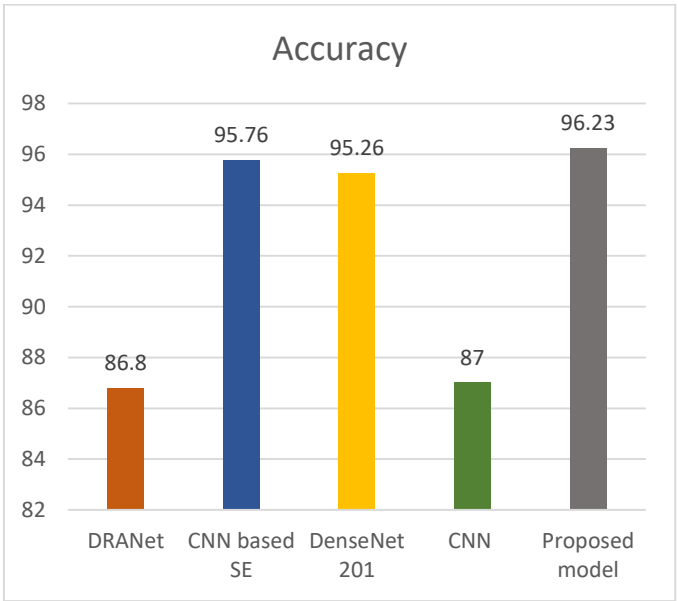

Fig. 10. Performance comparison.

The higher accuracy achieved by the ViT-based model is particularly noteworthy when considering the challenges associated with melanoma detection, including the complex and varied visual characteristics of skin lesions and the need for accurate differentiation between malignant and benign conditions. The model's ability to effectively handle these challenges, as evidenced by its outperformance of existing methods, highlights its robustness and adaptability. This advancement not only provides a powerful tool for clinicians but also addresses some of the key limitations of previous approaches, such as the extended training times, resource-intensive computations, and reduced efficacy in noisy or poor-quality images. The ViT-based model's success in overcoming these challenges and delivering high accuracy in skin cancer diagnosis positions it as a promising candidate for integration into clinical practice, where it could significantly improve the speed and accuracy of melanoma detection, ultimately saving lives through earlier and more precise intervention.

TABLE II. COMPARISON WITH EXISTING SYSTEM

Author	Methodology	Accuracy
Jiang et al	DRANet	86.8%
Shorfuzzaman et al	CNN based stacked ensemble framework	95.76%
Khan et al	DenseNet201 with transfer learning	95.26%.
Inthiyaz et al	CNN	87%
Proposed model	**Vi transformer based deep learning**	**96.23%**

VI. CONCLUSION

Melanoma, a form of skin cancer originating in melanocytes, presents a significant global health concern due to its aggressive nature and potential for metastasis. With its incidence steadily rising worldwide, melanoma detection and classification have become pivotal areas of research in the medical field. Early diagnosis plays a crucial role in improving patient outcomes, as timely intervention can significantly enhance treatment efficacy and prognosis. This research endeavours to develop a model capable of effectively detecting and classifying melanoma using publicly available datasets, with a focus on performance enhancement. The proposed approach introduces a Vision Transformer-based melanoma classification model adept at distinguishing between Benign and Malignant cases. With an accuracy of 96.23%, the model demonstrates a strong overall correctness in classification. These findings underscore the efficacy and potential of the proposed method in advancing the field of skin cancer diagnosis, offering promising prospects for enhanced accuracy and efficacy in clinical settings. These outcomes emphasize the effectiveness of the method in advancing skin cancer diagnosis, indicating promising prospects for heightened accuracy and efficacy in practical clinical settings. However, several avenues for future work could further advance the field. Future research could focus on expanding the dataset to include a more diverse range of skin types, lesion characteristics, and image quality conditions to improve the model's generalizability and robustness. Additionally, integrating the Vision Transformer with other advanced techniques, such as multi-modal data fusion or ensemble approaches, might enhance its performance further.

REFERENCES

[1] Parker, E. R., Mo, J., & Goodman, R. S. (2022). The dermatological manifestations of extreme weather events: a comprehensive review of skin disease and vulnerability. The Journal of Climate Change and Health, 8, 100162.

[2] Arnold, J. D., Yoon, S., & Kirkorian, A. Y. (2019). The national burden of inpatient dermatology in adults. Journal of the American Academy of Dermatology, 80(2), 425-432.

[3] Champsas, G., & Papadopoulos, O. (2020). The Role of the Sentinel Lymph Node Biopsy in the Treatment of Nonmelanoma Skin Cancer and Cutaneous Melanoma (pp. 647-704). Springer International Publishing.

[4] Saginala, K., Barsouk, A., Aluru, J. S., Rawla, P., & Barsouk, A. (2021). Epidemiology of melanoma. Medical sciences, 9(4), 63.

[5] Leong, S. P., Naxerova, K., Keller, L., Pantel, K., & Witte, M. (2022). Molecular mechanisms of cancer metastasis via the lymphatic versus the blood vessels. Clinical & Experimental Metastasis, 39(1), 159-179.

[6] Basurto-Lozada, P., Molina-Aguilar, C., Castaneda-Garcia, C., Vázquez-Cruz, M. E., Garcia-Salinas, O. I., Álvarez-Cano, A., ... & Robles-Espinoza, C. D. (2021). Acral lentiginous melanoma: Basic facts, biological characteristics and research perspectives of an understudied disease. Pigment cell & melanoma research, 34(1), 59-71.

[7] Zafar, K., Gilani, S. O., Waris, A., Ahmed, A., Jamil, M., Khan, M. N., & Sohail Kashif, A. (2020). Skin lesion segmentation from dermoscopic images using convolutional neural network. Sensors, 20(6), 1601.

[8] Adla, D., Reddy, G. V. R., Nayak, P., & Karuna, G. (2022). Deep learning-based computer aided diagnosis model for skin cancer detection and classification. Distributed and Parallel Databases, 40(4), 717-736.

[9] Shorfuzzaman, M. (2022). An explainable stacked ensemble of deep learning models for improved melanoma skin cancer detection. Multimedia Systems, 28(4), 1309-1323.

[10] Khan, M. A., Muhammad, K., Sharif, M., Akram, T., & de Albuquerque, V. H. C. (2021). Multi-class skin lesion detection and classification via teledermatology. IEEE Journal of Biomedical and Health Informatics, 25(12), 4267-4275.

[11] Jiang, S., Li, H., & Jin, Z. (2021). A visually interpretable deep learning framework for histopathological image-based skin cancer diagnosis. IEEE Journal of Biomedical and Health Informatics, 25(5), 1483-1494.

[12] Yacin Sikkandar, M., Alrasheadi, B. A., Prakash, N. B., Hemalakshmi, G. R., Mohanarathinam, A., & Shankar, K. (2021). Deep learning based an automated skin lesion segmentation and intelligent classification model. Journal of ambient intelligence and humanized computing, 12(3), 3245-3255.

[13] Zghal, N. S., & Derbel, N. (2020). Melanoma skin cancer detection based on image processing. Current Medical Imaging, 16(1), 50-58.

[14] Alwakid, G., Gouda, W., Humayun, M., & Sama, N. U. (2022, December). Melanoma detection using deep learning-based classifications. In Healthcare (Vol. 10, No. 12, p. 2481). MDPI.

[15] Cirrincione, G., Cannata, S., Cicceri, G., Prinzi, F., Currieri, T., Lovino, M., ... & Vitabile, S. (2023). Transformer-based approach to melanoma detection. Sensors, 23(12), 5677.

[16] Inthiyaz, S., Altahan, B. R., Ahammad, S. H., Rajesh, V., Kalangi, R. R., Smirani, L. K., ... & Rashed, A. N. Z. (2023). Skin disease detection using deep learning. Advances in Engineering Software, 175, 103361.

[17] Mao, X., Qi, G., Chen, Y., Li, X., Duan, R., Ye, S., ... & Xue, H. (2022). Towards robust vision transformer. In Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (pp. 12042-12051).

[18] Fanconi, C. (2019). Skin cancer: malignant vs. benign. *Distributed by ISIC Archieve*.