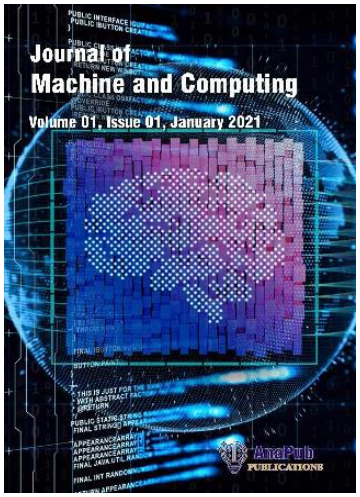
Journal Pre-proof

A Computing Framework for Transfer Learning and Ensemble Classification of Surface Patterns



Akepati Sankar Reddy and Gopinath M P

DOI: 10.53759/7669/jmc202505011

Reference: JMC202505011

Journal: Journal of Machine and Computing.

Received 02 May 2024

Revised form 16 September 2024

Accepted 16 October 2024

Please cite this article as: Akepati Sankar Reddy and Gopinath M P, “A Computing Framework for Transfer Learning and Ensemble Classification of Surface Patterns”, Journal of Machine and Computing. (2025). Doi: https:// doi.org/10.53759/7669/jmc202505011

AnaPub
PUBLICATIONS

A Computing Framework for Transfer Learning and Ensemble Classification of Surface Patterns

[1]Akepati Sankar Reddy, [2]Gopinath M.P
[1]Research Scholar, [2]Associate Professor,
[1,2] School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India
[1]akepatisankar.reddy2020@vitstudent.ac.in , [2]mpgopinath@vit.ac.in
Corresponding Author: Gopinath M.P

Abstract - The rapid increase in population density has posed significant challenges to medical sciences in the auto-detection of various diseases. Intelligent systems play a crucial role in assisting medical professionals with early disease detection and providing consistent treatment, ultimately reducing mortality rates. Skin-related diseases, particularly those that can become severe if not detected early, require timely identification to expedite diagnosis and improve patient outcomes. This paper proposes a transfer learning-based ensemble deep learning model for diagnosing dermatological conditions at an early stage. Data augmentation techniques were employed to increase the number of samples and create a diverse data pattern within the dataset. The study applied ResNet50, InceptionV3, and DenseNet121 transfer learning models, leading to the development of a weighted and average ensemble model. The system was trained and tested using the International Skin Imaging Collaboration (ISIC) dataset. The proposed ensemble model demonstrated superior performance, achieving 98.5% accuracy, 97.50% Kappa, 97.67% MCC (Matthews Correlation Coefficient), and 98.50% F1 score. The model outperformed existing state-of-the-art models in dermatological disease classification and provides valuable support to dermatologists and medical specialists in early disease detection. Compared to previous research, the proposed model offers high accuracy with lower computational complexity, addressing a significant challenge in the classification of skin-related diseases.

Keywords - Skin Cancer, Transfer Learning, Medical Image Processing, Ensemble Learning, Deep Learning, ISIC

I. INTRODUCTION

Skin cancer is a leading health concern globally, ranking as the second most common reason for death after heart disease. It is currently the 19th most typical cancer worldwide, with a concerning increase in incidences, and its rising incidence makes it a serious threat. This particular type of cancer grows when skin cells proliferate out of control, which results in the formation of malignancies. Melanoma and non-melanoma are the two primary types of skin cancers. Squamous cell carcinoma and basal cell carcinoma are the two most prevalent non-melanoma tumours. The primary cause of both melanoma and non-melanoma skin cancer is exposure to too much specific light, such as tanning beds or the sun's ultraviolet rays. Sunburn is associated with melanoma risk regardless of age: the more sunburned during childhood, adolescence, or young adulthood, the more significant the exposure. Our daily lives expose us to a variety of contaminants and hazardous substances that might affect our skin, which is the largest organ in the body. The skin is the body's outer organ that covers it and carries out several essential functions, including temperature regulation, organ protection, sweating out toxins to detoxify the body, fluid preservation, and organ protection. Environmental elements such as carcinogenic chemicals, the sun's UV radiation, and inflammatory agents are known to contribute to skin problems and cancers. Skin cancer has been growing worldwide due to increasing exposure to human-made pollutants and UV radiation. Scientists have found that a variety of environmental contaminants create reactive oxygen species (ROS), which can lead to oxidative stress, genetic damage, and abnormalities in our body's redox system. Though skin-related diseases are common, they can be difficult to diagnose and treat and have a big influence on a person's health. These conditions are directly related to an individual's general health and surroundings, including wool, particular foods, soaps, pet hair, and foods. These elements have the potential to activate the immune system and cause symptoms, including discomfort, redness, and itching. Skin problems can affect people of any age or gender. However, skin diseases continue because of problems with healthcare systems and a lack of resources to deal with these issues. It might be difficult to identify different types of skin cancer accurately, so specialists with knowledge in this field are needed [1]. The worrying findings around skin cancer under- score the importance of prevention and the need to take preventative measures. Common methods for diagnosing skin cancer include physical inspection, noninvasive dermoscopic, and biopsy of any concerning lesions. After the biopsy procedure, the malignant status of the sample was determined by microscopic analysis. If the biopsy confirms that cancer is present, further tests may be done to determine its exact type and extent. Depending on the type and stage of cancer, skin cancer treatment options may include surgery, radiation therapy, chemotherapy, immunotherapy, or other methods. The primary treatment option for early-stage skin cancer is surgical removal of the malignant lesion, which is frequently beneficial in treating the condition. Chemotherapy, immunotherapy, and radiation therapy are used to treat cancer that has spread to other anatomical areas, as well as metastatic cutaneous cancer [2].

Image processing is essential for accurately identifying medical images since it can extract important information that helps with the process. The process has multiple interrelated steps, which begin with improving the image quality using methods such as edge improvement, contrast enhancement, color correction, and lighting correction. Region growth, thresholding, and active contour modelling are some of the approaches used by image segmentation algorithms to split an image into regions of interest. Expert dermatologists obtain 50% to 60% accuracy rates in manual examinations [3], which indicate that more dependable techniques are required. Noise in dermoscopic images, such as hair, air bubbles, and other artefacts, in addition to lighting effects, can make skin cancer identification more challenging. It is essential to design an AI system which capable of autonomously identifying skin cancer from dermoscopic images. Pre-processing is essential for developing a system that removes artefacts like hair and increases image quality. Scientists have created many pre-processing methods, such as image inpainting, to match the values of nearby hair pixels. These methods are based on morphological operations and contrast enhancement. The ABCD rule and the seven-point checklist are two common tools used to identify features in dermoscopic images. The ABCD rule considers four factors: the skin lesion diameter, color change, border irregularity, and asymmetry.

The detection and classification of skin cancer have benefited greatly in recent years by applying the techniques of deep learning (DL) and machine learning (ML) [4]. Algorithms such as decision trees, support vector machines (SVMs), and DL are essential to dermatology because they can analyze large amounts of data that include images of the skin. After all necessary steps have been completed, ML and DL models are used to classify the final output accurately. These methods work together to provide an accurate and trustworthy cancer classification. ML and DL techniques are superior to older methods in a number of ways when it comes to the detection and classification of skin cancer. Although the majority of the work was exceptional in accuracy, there are some overfitting issues, and the model might be working well for particular classes. We are making an effort to fill those gaps. As skin cancer becomes more common worldwide, research is concentrated on enhancing the early diagnosis of skin diseases. This drives the goal of providing medical professionals with better tools so they can diagnose patients earlier and possibly save lives. The pressing need to increase diagnostic accuracy especially in the case of skin cancer, where early diagnosis greatly enhances patient outcomes motivates this research. The study's principal contribution is as follows:

- We have designed an Ensemble Deep-Learning model for classifying skin cancer.
- We employed data augmentation techniques to enhance the dataset size and improve generalization, robustness to variability, balance classes, and reduction of annotation.
- We executed several experiments, including operating different combinations of base learners, which improved the overall accuracy.
- We discovered that skin cancer classification could be efficiently performed by ensemble learning and transfer learning models.
- We fine-tuned the transfer learning models as base learners.
- Our proposed model not only enhances accuracy but also significantly streamlines computational processes, paving the way for seamless deployment.

This paper has been organized into multiple sections. First, we will review the existing methods for identifying skin cancer in Section II. Then, in Section III, we will provide an overview of our methodological statement. In Section IV, we briefly discuss the experimental setup for our model implementation. The experimental results of the models we have employed will be presented in Section V. Ultimately, we will showcase our research findings in Section VI

II. LITERATURE REVIEW

Over the years, a number of researchers have conducted numerous studies to improve the efficacy in the field of classifying skin cancer. Here are a few noteworthy and current works on this particular area of skin cancer research.

Nasr-Esfahani et al. [5] proposed SNC Net, A novel technique to automatically identify eight different forms of skin cancer from dermoscopic images. They combined DL models and handcrafted (HC) feature extraction to improve classification accuracy. SNC Net, which was trained on the ISIC 2019 dataset, achieved an accuracy of 97.81%, precision of 98.31%, recall of 97.89%, and F1 score of 98.10%, outperforming four baseline models and six state-of-the-art (SOTA) classifiers. Monica et al. [6] presented an automated framework to improve the early detection of Melanoma Skin Cancer (MSC) by handling challenges in analyzing dermoscopic images. They incorporated normalization techniques, precise segmentation employing mask-faster RCNN, and feature extraction with pre-trained CNN models. The modified GRU model, combining swish-ReLU activation, performed exceptional accuracies of 99.95% and 99.98% on ISIC 2020 and HAM 10000 datasets, respectively, outperforming conventional methods.

Campos et al. [7] proposed a unique treatment approach for non-melanoma skin cancer (NMSC) employing a variety of photothermal and chemotherapeutic techniques. MoS2 nanoparticles are used as a photothermal agent, and these nanoparticles are synthesized utilizing a liquid-phase exfoliation and intercalation technique with polyvinylpyrrolidone (PVP), performing in an average particle size of 165 ± 170 nm after ultrasonication treatment. Experimental outcomes confirmed that MoS2 nanosheets efficiently convert near-infrared (NIR) light into heat, acquiring temperatures of 52°C. MoS2 (125 μg/mL) and Tegafur (50 μg/mL) therapeutic dosages were combined and adjusted into a Carbopol hydrogel. Metabolic activity tests on normal human fibroblasts (HFF-1 cells) showed that they did not drop below the 70% toxicity threshold. However, metabolic activity dramatically dropped in skin cancer cells A-431. Notably, the integrated MoS2 + Tegafur hydrogels lead to a 1.9-fold reduction in A-431 cancer cell metabolic activity 72 hours after irradiation corresponding to MoS2 alone.

Rahman et al. [2] presented an optimized DCNN-based model for accurately classifying skin cancer as malignant or benign melanoma employing dermoscopic images. Adding more data and an additional basic layer to the NASNet architecture to manage inconsistent and partial data is where the uniqueness lies. The proposed approach produced encouraging outcomes with a dataset of 2637 skin image captures. Precision, sensitivity, specificity, F1-score, and area under the ROC curve were examples of evaluation measures. After optimization, the Adam optimizer generated 85.6% and 83.7% accuracy for the NASNet Mobile and NASNet Large models, respectively. Naeem et al. [8] introduced DVFNet, a DL-based approach for detecting skin cancer from dermoscopy images. DVFNet coalesced VGG19 architecture and Histogram of Oriented Gradients (HOG) for discriminative feature extraction, improving accuracy. Pre-processing with anisotropic diffusion enhanced image quality, while SMOTE Tomek addresses imbalanced image classes. Damaged skin cell locations can be identified with the use of segmentation, and using HOG and VGG19 features, a feature vector map is generated. CNN classification gained 98.32% accuracy on the ISIC 2019 dataset and ANOVAstatistical test validated model accuracy.

AlSadhan et al. [9] investigated the significance of four unified convolutional neural networks (YOLOv3, YOLOv4, YOLOv5, and YOLOv7) in classifying skin lesions for earlier and proper diagnosis of skin cancer. Each of the models was tested in terms of lesion localization, classification accuracy, and inference time using a benchmark dataset. Notably, YOLOv7 surpassed the other models with an IoU of 86.3%, mAP of 81.4%, F1-measure of 80%, and an inference time of 0.32 seconds per image. Kousis et al. [10] addressed the continuing challenge of properly recognizing skin cancer from medical images, con- centrating on developing a mobile application employing a single DL model. Then, 11 CNN architectures were trained and tested on the HAM10000 dataset, using techniques such as data augmentation, transfer learning, and fine-tuning to embed imbalances and likeness between lesions. DenseNet169 appeared as the top-performing model, performing an accuracy of 92.25%, recall of 93.59%, and F1-score of 93.27%. A light version of DenseNet169 was incorporated into a mobile application, permitting users to classify lesions as benign or malignant operating their device's camera. Furthermore, based on UV radiation, skin phototype, and sunscreen use, the app offered information on safe sun exposure.

Rajesh et al. [11] used the HAM10000 dataset to train Deep Convolutional Neural Networks (DCNNs), which were optimized from the successful ImageNet dataset for classifying seven different types of skin lesions. They concentrated on Inception V3 and DenseNet 201 architectures. Inception V3 achieved 85.94% accuracy, DenseNet 201 reached 87.42% accuracy, and an ensemble model incorporating both performed 85.94% accuracy. In the validation, DenseNet 201 performed better in the test set with an accuracy rate of 87.42%, while Inception V3 just slightly surpassed DenseNet 201 over 20 epochs. Sivakumar et al. [12] introduced an automated diagnostic model and web application for identifying Malignant Melanoma Cancer, discoursing the demand for more efficient and precise diagnostic mechanisms for skin diseases. The used a convolutional neural network (CNN) with ResNet50 for data collection, preprocessing, segmentation, feature extraction, and classification. Novel preprocessing processes and mixed hybrid pooling phases improved accuracy by eradicating noise and enhancing spectral image information. Performance analysis demonstrated the model performed with 94% accuracy and an F1-score of 93.9% on the ISIC dataset, presenting noteworthy advancement over traditional methods. Furthermore, in comparison to current methods, the corresponding online application expedited diagnosis with accurate information, automating the categorization process and lowering the possibility of misdiagnosis.

After the comprehensive research, we identified specific constraints connected to skin cancer –

- The existing datasets have some drawbacks, such as an excess of tiny lesions, which make it challenging to train the algorithms due to limited diversity and high-class imbalance.
- Although advancement has been made in the field, the datasets still need to be scaled up through comprehensive data augmentation and analysis of improved imaging data.
- While algorithms have individual strengths, the challenge of optimization and generalization poses a considerable gap in their applicability. Furthermore, the computational complexity of the deep learning model is a major challenge.

- Significant accuracy limits exist when extracting features from skin lesions, which creates an extensive gap in accuracy and reliability.
- Despite progress, frameworks still need work to acquire high efficiency in terms of time and memory usage, contextual understanding, and trade-off between accuracy and speed. These gaps represent noteworthy limitations in the frameworks' computational efficiency and scalability.

III. MATERIALS AND METHODS

In this paper, we present a unique method for classifying skin image data employing an ensemble approach, as demonstrated in Figure 1. This method can be extended to enclose any group of models within an ensemble. The next sections go into further detail about our innovative approach.

3.1 Base Learners

The procedure of selecting base learners can vary depending on the difficulty at hand. Still, the primary objective is to choose the most appropriate models for the assigned problem. For our classification task, we opted for three distinctive pre-trained CNN models: ResNet50, InceptionV3, and DenseNet121. DenseNet-121 is a DenseNet variant that uses densely connected blocks to improve information flow throughout the network, making it suitable for tasks like image classification on the ImageNet dataset. DenseNet improves gradient flow and feature reuse by concatenating feature maps instead of summing them, in contrast to conventional feed-forward neural networks. Richer feature representations are fostered by this concatenation mechanism, which guarantees that each layer has direct access to the gradients from all previous layers. The model is composed of Dense Blocks, where the growth rate hyperparameter determines how many feature maps each layer within a block adds to the collective knowledge. Transition Blocks employ 1x1 convolutions, 2x2 pooling layers, and batch normalization to enable downsampling and dimensionality reduction. DenseNet-121 divides the volume and feature maps in half after each Transition Block, but the volume inside a Dense Block stays constant. A 1x1 convolution shrinks the size of each feature map inside a Dense Layer, and a 3x3 convolution with the number of feature maps growing at a constant pace comes next. The volume of input and the output of these operations are then concatenated, thereby continuously improving the network's collective knowledge. To achieve state-of-the-art performance in image classification tasks, the DenseNet-121 model uses dense connectivity to leverage iteratively improve its representations. Microsoft Research created ResNet50, a deep convolutional neural network architecture that is well-known for its remarkable performance in image classification tasks, in 2015. It has 50 layers and adds residual connections so that deeper networks can be trained without worrying about vanishing gradient problems. Convolutional layers, identity blocks, convolutional blocks, and fully connected layers are the four primary parts of the architecture. Convolution, batch normalization, and ReLU activation are the methods used by convolutional layers to extract features from input images. Max pooling is then used to reduce spatial dimensions. Convolutional neural network (CNN) architecture InceptionV3, which is optimized for image classification, is a major advancement over the Inception model, which is its predecessor. Finding a balance between classification accuracy and computational efficiency is its main objective.

The core idea behind InceptionV3 is "Inception modules", which are essential parts that include max-pooling operations and convolutional filters with different sizes (1x1, 3x3, and 5x5). The network can efficiently capture features at various spatial scales in a single layer thanks to this amalgamation. Furthermore, InceptionV3 employs a number of architectural techniques designed to reduce the number of parameters and computational overhead without sacrificing the quality of its representations. For example, dimensionality reduction is achieved through the widespread use of 1x1 convolutions, which reduces computational complexity. Furthermore, huge convolutions are divided into smaller ones using methods like factorization, which reduces the number of parameters even more. These models are broadly acknowledged CNN architectures comprehended for their state-of-the-art performance. Each of these models includes exceptional structural properties, giving them different abilities to generalize the given allocation.

3.2 Weighted Ensemble Model

After every model has been selected, each is trained independently on the dataset, and their corresponding accuracy percentages, which are acc_x recorded on a validation set. It is notable that during training, this validation set is not used. The models learn the same underlying distribution differently due to their heterogeneous nature and use of multiple hyper-parameters. These models are trained continuously until they reach convergence or until there is no further decrease in the loss value. Subsequently, after training, the models undergo evaluation using a validation dataset, where their accuracies are measured and recorded. These recorded accuracies are then employed to estimate the weight factor α_x for every respective model, utilizing Equations 1 and 2. Here, x varies from 1 to a, where 'a' represents the entire number of models.

In Equation 2, we increment 1 to αx because 'β' describes the accuracy of the least accurate model. It secures that the weight α_x for the least accurate model equals 1 by adding 1 when it is inserted into 1. The least accurate model might be excluded if this adjustment was made because it would not have a substantial impact on the ensemble model's final result. These recorded accuracies are then used to calculate the weight factor α_x for each respective model employing the following equations:

$$a_x = acc_x - \beta + 1, \text{x=1} \ldots\ldots\ldots\ldots \text{a} \quad (1)$$

$$\beta = \min([acc_x]_{x=1}^{a}) \quad (2)$$

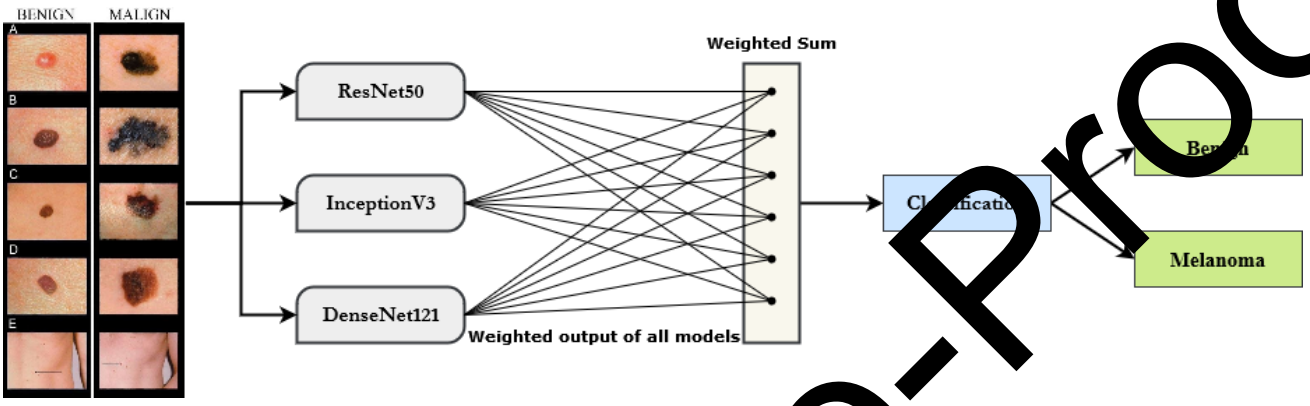


Fig. 1. Proposed weighted ensemble model for skin cancer classification

1) Aggregating of Ensemble Model: Each model is used to make predictions for input data after it has been trained. After that, a SoftMax function is applied to these predictions, the output of the model for every input sample K_p yielding probability vectors denoted by C_g , where g is the number of classes, ranging from 1 to n_c. After then, the output of each model, $[M_x]_{x=1}^{a}$, is multiplied by the corresponding multiplication factor, α_x. Equation 3 is used to add the weighted probability from each model to generate the output of the ensemble model for a particular class. Algorithm 1 describes the weighted ensemble model algorithm, which reduces the complexity of understanding and implementing. Figure 1 illustrates the diverse multiplying factors (α_x) for each model, characterized by different colored lines that are specified by individual accuracies when considering a specific model, αx stays constant, and Equation 3 is employed to specify the output for a particular class. The equation for calculating the ensemble model's output for a particular class T_k is:

$$\text{OutputClass } K_{p,ensemble} = max_{x=1}^{a} \left(\sum_{k=1}^{nc} a_x \; x \; [C_g]k, x\right) \quad (3)$$

Algorithm 1 delineates the steps for our presented weighted ensemble model. First of all, we partitioned our dataset into three distinct parts in order to prevent it: training, validation, and testing. We load the initial portion, Let D = {(m1, n1), . . ., (mh, nh)} where m_k is the set of feature vectors and y_i the connected labels, as the training dataset D. The residual amount is divided into two groups: the testing set T and the validation set V. The number of models, denoted by M (in our case, M = 3), we utilize every model to produce predictions on the validation dataset $[V_k]_{k=1}^{g}$, where k = 1, 2, . . ., g. Once trained, each model make predictions on the validation dataset V, producing predictions V_k for i= 1 to m. The accuracy of each model is then calculated using these predictions, which are represented as M_x. We then compute a parameter β to help determine αx the weight factor for each model. Next, the predictions of every model are multiplied by this weight factor, α_x. After combining the weighted predictions from each model, the proposed weighted ensemble model (WEM) is generated. In order to use this output for prediction, we first use the argmax function to find the class that the WEM has predicted for a particular sample. An alternative way to get the same result is to use a SoftMax function to determine the class from the WEM's output.

3.3 Weighted Ensemble Model

The most popular and straightforward strategy is the model averaging ensemble method. This method uses an average of the base learners' output to determine the ensemble model's final prediction. Merely averaging the ensemble models enhances the generalization performance by reducing the variance among the models, which is a result of deep learning architectures' high variance and low bias. This is due to deep learning models' propensity to overfit, which results in lower

validation accuracy and higher training accuracy. Deep learning models consequently struggle to generalize to new data. By averaging the predictions from the various base learners, this is prevented. The variance between the models is decreased by averaging the predictions, which results in accurate generalization performance.

Algorithm 1:

Ensemble Model Input:

1. D = {(m1, n1), . . . , (mh, nh)} ▷ Training dataset
2. V = {(m1, n1), . . . , (mg, ng)} ▷ Validation dataset
3. T = {(m1, n1), . . . , (mu, nu)} ▷ Testing dataset
4. **for** x = 1 **to** m do ▷ Iterate over models
5. Mx ▷ xth model
6. Classify data into n_c classes
7. **end for** = 0

Ensemble Model Output Calculation:

1. **Output:** C_g is a vector of probabilities, k = 1 . . . n_c, OutputClassKp, M is a vector defining the output class of Kp employing model M.
2. D, dataset; m, number of models; g, number of classes; nc, number of samples in ; V_k, input vector.
3. **for** x = 1 to m **do**
4. Train model M_x with dataset D.
5. Calculate the class probabilities for V_k, k = 1 . . . g employing model M_x.
6. $[C_g]k,x = M_x(V_k)$.
7. Calculate the output class of Vk for k = 1 to m using $[C_g]k,x$.
8. Output$ClassV_k$, M_x = max$[C_g]k,x$.
9. Calculate the accuracy accx of model Mx on V employing Output lass , Mx(=1) g against True-Label.
10. **end for**
11. Calculate parameter β = min($[acc_x]_{x=1}^{a}$).
12. **for** x = 1 to m **do**
13. Estimate weight of model Mx: $\alpha_x = (acc_x) - \beta + 1$
14. **end for**
15. Calculate the output of the proposed model for Kp, k = 1 . . . n.
16. OutputClassKp, ensemble = max $\sum_{x=1}^{a=1} \sum^{nc} a_x\, x[(C_g)]k,x)$.
17. Calculate the accuracy $acc_{ensemble}$ utilizing Out ClassK$_{p,ensemble}$ $_{k=1}^{n}$ against True-Label.

Either the predicted probabilities of the classes are averaged using the SoftMax function, or the outputs of the base learners are directly averaged. In situations where the performance of the base learners is comparable, unweighted averaging makes sense. It is possible to improve overall performance because some learners may perform poorly overall but perform exceptionally well when classifying particular subclasses. The adaptive meta- learner should be able to combine the strengths of the base learners in an adaptive manner. Additionally, to illustrate the variance reduction through averaging, the variance of the ensemble prediction Var[prediction] can be calculated as:

$$\text{Var[Predict}\ ] = \frac{1}{N^2} \sum_{i=1}^{N} Var\,[pi] \;+\; \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} Cov\,[pi, pj]\ \frac{1}{N^2} \qquad (4)$$

where Var[p_i] denotes the variance of the probability vector output by model i and Cov[p_i, p_j] denotes the covariance between the probability vectors output by models i and j.The architecture of the average ensemble model is shown in Figure 2. The hyperparameters and their values of the ensemble models are shown in Table I.

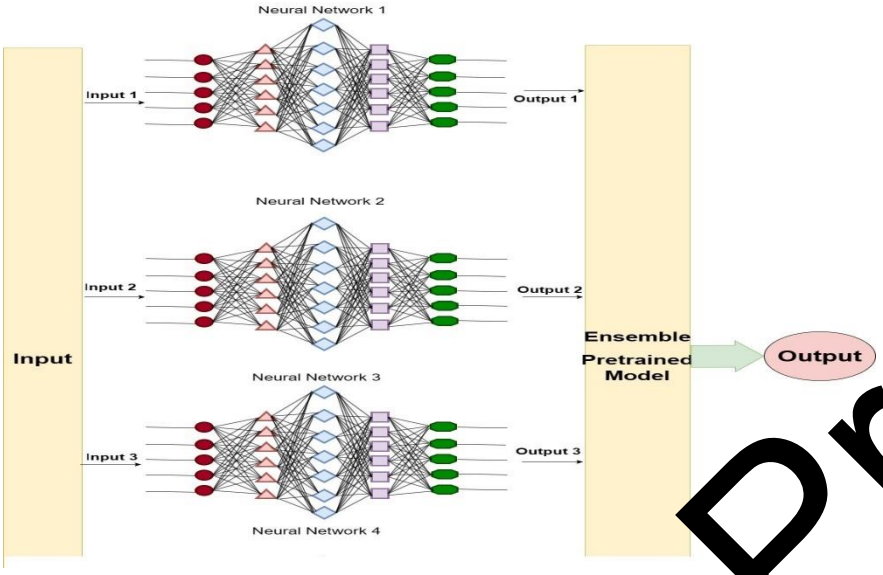


Fig. 2. Average ensemble model for skin cancer classification

TABLE I. HYPERPARAMETERS AND VALUES

Hyperparameters	Value
Base learners	Transfer learning
Batch size	1
Epochs	15
Input image	299*299
Hidden activation	Relu
Output activation	SoftMax
Loss	Categorical Cross entropy
Optimizer	Adam
Trainable parameters	5022805

IV. EXPERIMENTAL SETUP

4.1 Implementation & Hardware Specifications

We used hardware resources for our model, which included one Tesla K80 GPU with 2496 CUDA cores and 12 GB of GDDR5 VRAM. Significant processing power was made available by this GPU, which made training and inference procedures more effective. In addition to the GPU, our system included a 2.3 GHz Xeon CPU with 4 cores and 8 threads to facilitate efficient completion of tasks involving computationally demanding tasks. We also made use of the 12 GB of RAM that was available to store and manipulate data for model training and assessment. In terms of storage, our system managed model files and datasets with 2 GB of disk space. A wide range of tools improved our model implementation in terms of software and libraries. We used numpy for effective numerical operations, scikit-learn for a variety of tasks, and TensorFlow as a deep learning framework, aided by the Keras high-level API for streamlined model development. Pandas also made preprocessing and data manipulation easier, which improved the productivity of our workflow. Together, these software elements enabled scalable and reliable performance during the development and implementation of our model.

4.2 Dataset Description & Preprocessing

The International Skin Imaging Collaboration (ISIC) dataset, which includes 2357 pictures of benign and malignant oncological conditions, was used in our study. Except for melanomas and moles, whose images are somewhat predominant, all images were sorted in accordance with the ISIC classification, and each subset was split into the same number of images. Actinic keratosis, basal cell carcinoma, dermatofibroma, melanoma, nevus, pigmented benign keratosis, seborrheic keratosis, squamous cell carcinoma, and vascular lesion are among the diseases included in the data set. Next, we divided the dataset into two categories: Melanoma and Benign. In Figure 3, the dataset sample is displayed. We have applied

several preprocessing steps in our research. First of all, we have applied the data augmentation technique in our research to increase the data samples. One popular technique in computer vision tasks is data augmentation, which is creating new training examples by transforming existing data samples in different ways. There are two main goals for data augmentation. First of all, it artificially increases the size of the dataset, which is advantageous when working with small amounts of data. More diverse examples are available to models during training when the dataset size is increased, which improves performance and robustness, especially for deep learning models. Secondly, by adding noise and variability to the training set, data augmentation functions as a type of regularization. This helps avoid overfitting, a phenomenon in which models fail to learn generalizable patterns and instead memorize the training set. In order to guarantee that there is a sufficient balance between the classes in our dataset, data augmentation is also applied to it prior to training. When compared to a class with fewer training images, the one with more training images will be biased to obtain higher accuracy. The images were enhanced using an image data generator from Keras preprocessing integrated with TensorFlow. Table II lists the augmentation's parameters.

Fig. 3. Sample of the dataset

TABLE II. AUGMENTATION TECHNIQUES OF TRAINING IMAGES

Process Name	Value
Rescale	1/255
Rotation	10
Width Shift	0.2
Height Shift	0.2
Shear	0.2
Horizontal Flip	True
Vertical Flip	True
Fill Mode	nearest

Pixel Standardization: Using the well-known Gaussian pixel-standardization method, proper data scaling is also guaranteed prior to feeding the training images into the model. Equation 5 describes this technique, which involves dividing the result by the standard deviation of the pixel values in a training image and subtracting the mean pixel value from each pixel.

$$X = \frac{X - \mu pixel}{\sigma pixel} \quad (5)$$

where X is the original variable, μpixel is the mean of the pixel values, and σpixel is the standard deviation of the pixel values. We used a tripartite data division strategy with training, testing, and validation datasets very carefully in our research. This approach is essential to guaranteeing our models' generalizability, robustness, and dependability. To be more precise, we divided the dataset into 500 for validation and 500 for testing. This strategy has a complex justification. First off, model learning and parameter optimization are based on the training dataset. Our models can better identify underlying patterns and features and perform better in prediction when they are exposed to a wider variety of samples. To further strengthen the model's ability to generalize to new data and reduce overfitting tendencies, the training dataset can be expanded. On the other hand, the 500-image testing dataset serves as a neutral benchmark for assessing the model's performance. We are able to evaluate the extent to which our models generalize to new, unseen data by withholding these samples during training. This guarantees that the performance metrics acquired are not distorted by overfitting and are

representative of the model's actual efficacy. In addition, the validation dataset—which consists of an additional 500images—is essential for optimizing model hyperparameters and guarding against data leaks. To effectively optimize our models without tainting the testing data, we can make iterative adjustments to the model configurations based on performance metrics on the validation set. Class-wise training, testing and validation samples are shown in Table III.

TABLE III. CLASS-WISE TRAINING, TESTING AND VALIDATION SAMPLES

Evaluation Metrics	Training	Testing	Validation
Benign	2000	250	250
Melanoma	2000	250	250

4.3 Evaluation Parameters

In the final phase of the suggested ensemble-based model, the aggregated prediction values for each class in each model are used to predict the lesions class. The accuracy of a model's class prediction is used to assess classification performance. All of the widely used performance metrics—accuracy, precision, recall, f-score, and AUC—are used in this evaluation to support the model's high performance. These metrics are defined below.

• **Precision**: Precision measures the accuracy of positive predictions. It is the ratio of true positive predictions to the total predicted positives.

$$\text{Precision} = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

• **Recall (Sensitivity):** Recall measures the proportion of actual positives that were correctly identified by the model. It is the ratio of true positive predictions to the total actual positives.

$$\text{Recall} = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

• **F-score (F1-score):** The F-score is the harmonic mean of precision and recall. It balances both measures and provides a single score that considers both false positives and false negatives.

$$\text{F 1-score} = \frac{2 \times Precision\ x\ Recall}{Precision + Recall}$$

• **Accuracy:** Accuracy measures the overall correctness of the model. It is the ratio of correct predictions to the total number of predictions.

$$\text{Accuracy} = \frac{True\ Positives + True\ Negatives}{Total\ Predictions}$$

• **Kappa (Cohen's Kappa):** Kappa statistic measures the agreement between the predicted and actual classifications while accounting for the possibility of the agreement occurring by chance.

$$\text{K} = \frac{Po - Pe}{1 - Pe}$$

where p_o is the relative observed agreement, and p_e is the hypothetical probability of chance agreement.

• **Matthews Correlation Coefficient (MCC):** The MCC takes into account true and false positives and negatives and is generally regarded as a balanced measure that can be used even if the classes are of very different sizes.

• **Confusion Matrix:** A confusion matrix is a table used to describe the performance of a classification model. It presents the actual and predicted classes in a tabular format, showing correct and incorrect predictions.

• **Area under the ROC Curve (AUC):** AUC measures the area under the receiver operating characteristic (ROC) curve. It quantifies the model's ability to discriminate between positive and negative classes across different thresholds.

$$\text{AU C} = \int_0^1 TPR\ (FPR^{-1}(t))dt$$

Where TPR is the true positive rate, and FPR is the false positive rate.

V. RESULTS AND DISCUSSION

In this section, we will briefly explain our model's outcomes on the given datasets and compare them with those of other models that used the same dataset. The performance of ResNet50, InceptionV3, and DenseNet121 on the ISIC dataset demonstrated variations in testing, training, and validation accuracy.

5.1 Outcomes of Our Innovative Model Implementation

Initially, ResNet50 performed a testing accuracy of 85%, with training and validation accuracies both at 75%, and this indicated some overfitting, as the model achieved better results and the training data than on unseen testing or validation data. Overfitting occurs when a model comprehends to memorize the training data rather than generalize patterns. Move on to, InceptionV3 performed better, with a testing accuracy of 92%, training accuracy of 96%, and validation accuracy of 82%. While the training accuracy was heightened, the testing and validation accuracies were relatively down, suggesting some degree of overfitting. Then, DenseNet121 stood out with increased accuracies across all sets: testing at 98%, training at 98%, and validation at 88%. This indicated that the model learns representations that generalize well to unseen data, underestimating overfitting. Finally, when evaluating ensembles, both weighted and average, they display enhanced testing accuracies corresponding to individual models. The weighted ensemble performed the most increased testing accuracy of 98.5%, with training and validation accuracies also at 98.5% and 89%, respectively. The average ensemble, while not as higher as the weighted ensemble, still exceeded individual models with a testing accuracy of 94%, training accuracy of 97%, and validation accuracy of 92%. In Table IV, we presented the proposed model's performance in terms of testing, training, and validation accuracy.

TABLE IV. THE PROPOSED MODEL'S PERFORMANCE IN TERMS OF TESTING, TRAINING, AND VALIDATION ACCURACY

Model	Testing Accuracy	Training Accuracy	Validation Accuracy
ResNet50	85%	75%	75%
InceptionV3	92%	96%	82%
DenseNet121	98%	98%	88%
Weighted Ensemble	98.5%	98.5%	89%
Average Ensemble	94%	97%	92%

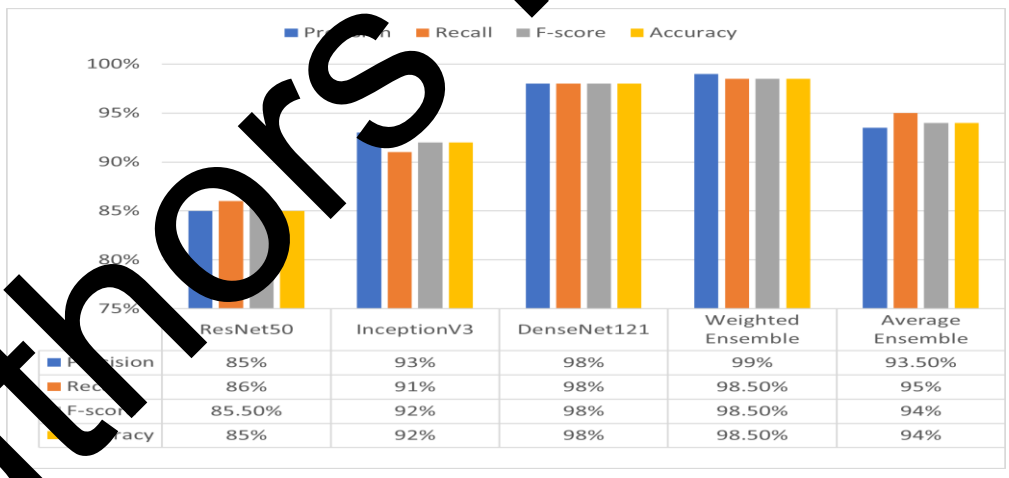


Fig. 4. The evaluation metrics of the models

The way that models were trained is the reason for differences in testing, training, and validation accuracy. To minimize the loss of the training data, the model modified its parameters during training. Nevertheless, if the model becomes very specialized to the training set and needs help to generalize to new data, this optimization may result in overfitting. A useful tool for fine-tuning hyperparameters and preventing overfitting during training was the validation set, which served as a stand-in for unavailable data. A model's final capacity to generalize was evaluated by testing accuracy, which is determined on an entirely independent dataset. Although each model has strengths, DenseNet121 was the most reliable since it regularly achieved high accuracies on all sets. Ensembles, particularly the weighted ensemble, increased performance even further, proving that it is effective to integrate different models to enhance overall accuracy and reduce shortcomings.

Among the models assessed, DenseNet121 identified both true positives and false positives with a score of 98%, which demonstrated its precision and recall. DenseNet121 exhibited consistent and dependable performance throughout the dataset, as evidenced by its 98% accuracy and F-score, which highlighted the balance between precision and recall. Comparatively, ResNet50 and InceptionV3 also performed well but less impressively than DenseNet121. InceptionV3 showed slightly higher precision and recall than ResNet50, resulting in a slightly improved accuracy and F-score. However, the Weighted Ensemble model outperformed individual models, reaching a precision of 99% and a recall of 98.50%. The ensemble model combined the strengths of different architectures, resulting in enhanced overall performance. On the other hand, the Average Ensemble, while acting decently with a 94% accuracy, was marginally inferior to the Weighted Ensemble. It achieved a precision of 93.50% and a recall of 95%, suggesting that it may need to be more practical in capturing all appropriate instances as the Weighted Ensemble. In Figure 4, we showed the evaluation metrics of the models in terms of Precision, Recall, F-score and Accuracy. Although every model performs admirably, the Weighted Ensemble performs better than others, demonstrating the value of mixing several models for optimal results.

In evaluating our ensemble models for the class imbalance dataset, we relied on two robust metrics: Kappa and MCC (Matthews Correlation Coefficient). These metrics present subtle insights into classification performance beyond uncomplicated accuracy, which is crucial for understanding the effectiveness of our models. The Weighted Ensemble Model presented outstanding agreement between predicted and actual classes, which is evidenced by its high Kappa score of 97.50%. Furthermore, its MCC value of 97.67% indicated the model's powerful binary classification quality regarding the intricacies of true and false positives and negatives, specifically relevant for imbalanced datasets. These metrics not only deliver an exhaustive evaluation of our proposed ensemble models but also enable informed comparisons between various approaches, assisting in the selection of the most useful solution for our specific problem domain. In the table V, we represented Kappa and MCC values for our two ensemble models.

TABLE V. AUGMENTATION TECHNIQUES OF TRAINING IMAGES

Model	Kappa	MCC Value
Weighted Ensemble Model	97.50%	97.67%
Average Ensemble Model	93.19%	93.65%

The comparison of our model with other existing work ensemble models, which are trained on the same dataset, is shown in Table VI. We are comparing our model with the other models that haven also used the same dataset. Our proposed model surpassed existing ensemble models by a considerable margin, boasting an outstanding accuracy of 98.50%. While the existing ensemble models demonstrated admirable accuracies ranging from 93% to 97.1%, our model surpassed them with a significant improvement. Moreover, our model performed this heightened accuracy while maintaining strong performance metrics across different evaluation criteria. In essence, the proposed model significantly improves predictive accuracy compared to specified ensemble methods, making it a convincing choice for applications demanding accurate and trustworthy predictions.

TABLE VI. COMPARING EXISTING MODELS WHICH APPLIED TO THE ISIC DATASET WITH OUR PROPOSED ENSEMBLE MODEL

Reference	Model	Accuracy
[13]	Ensemble	93%
[14]	Ensemble	97.1%
[15]	Ensemble	95.76%
[16]	Ensemble	97% (ROC-AUC)
Proposed Model	Ensemble	98.50%

In this study, we evaluated our proposed model against existing works that employed the ISIC dataset. We also evaluated our model's performance on additional datasets to ensure it was robust and broadly applicable across various data sources. This approach not only strengthens the validity of our results but also shows how our model can be used in situations outside of the initial training dataset.

In comparing existing models that applied to other datasets, it's crucial to consider the characteristics of the datasets they were trained on, which are depicted in Table VII. The HAM10000 dataset is a famous benchmark in dermatology, including a diverse range of skin lesion images. Models trained on this dataset, such as Efficient Nets and the Deep Learning-Based model, presented solid performances, with accuracies surpassing 87.91%. However, the accuracy of the Deep Learning-

Based model slightly outperformed the Efficient Nets with 91% accuracy, demonstrating the effectiveness of its architecture or training procedure. On the other hand, the Biomedical Datasets showed slightly lower accuracy compared to HAM10000, exhibiting potential differences in image quality or diversity. However, at 87.42%, the performance is still commendable. Deep-transfer learning provided an outstanding 98.61% accuracy on the PH2 dataset, which is renowned for its high-quality images and precisely annotated lesions. This represented a considerable improvement in accuracy and highlighted the effectiveness of employing transfer learning techniques on specialized datasets. Lastly, our proposed ISIC Ensemble model demonstrated comparable performance to the PH2 dataset, performing an accuracy of 98.50%. This suggested that the ensemble approach, likely leveraging various models or data sources, can effectively capture and classify dermatological features. Although every dataset has its unique features and difficulties, models developed using ensemble techniques and trained on datasets such as PH2 demonstrate encouraging results in pushing the boundaries of dermatological image categorization accuracy.

TABLE VII. COMPARING EXISTING MODELS WHICH APPLIED TO THE OTHER DATASETS WITH OUR PROPOSED ENSEMBLE MODEL

Reference	Dataset	Model	Accuracy
[17]	HAM10000	Efficient Nets	87.91%
[18]	HAM10000	Deep Learning-Based	91%
[19]	Biomedical datasets	Deep Learning-Based	87.42%
[20]	ph2 dataset	Deep Learning-Based	98.61%
Proposed Model	ISIC	Ensemble	98.50%

We carefully evaluated our weighted ensemble model's performance using both quantitative analysis and graphical evaluation. We were able to observe the model's capacity to generalize to new data and obtain insights into its learning dynamics by charting the training and validation accuracies over different epochs. In addition to giving us a thorough understanding of the training procedure, this visualization made it easier to spot any possible overfitting or underfitting. The epoch-wise evolution of training and validation accuracy for our proposed model is elegantly shown in Figure 5. This graphical representation gives important insights into the model's performance dynamics during training in addition to providing a visual representation of the learning trajectory.

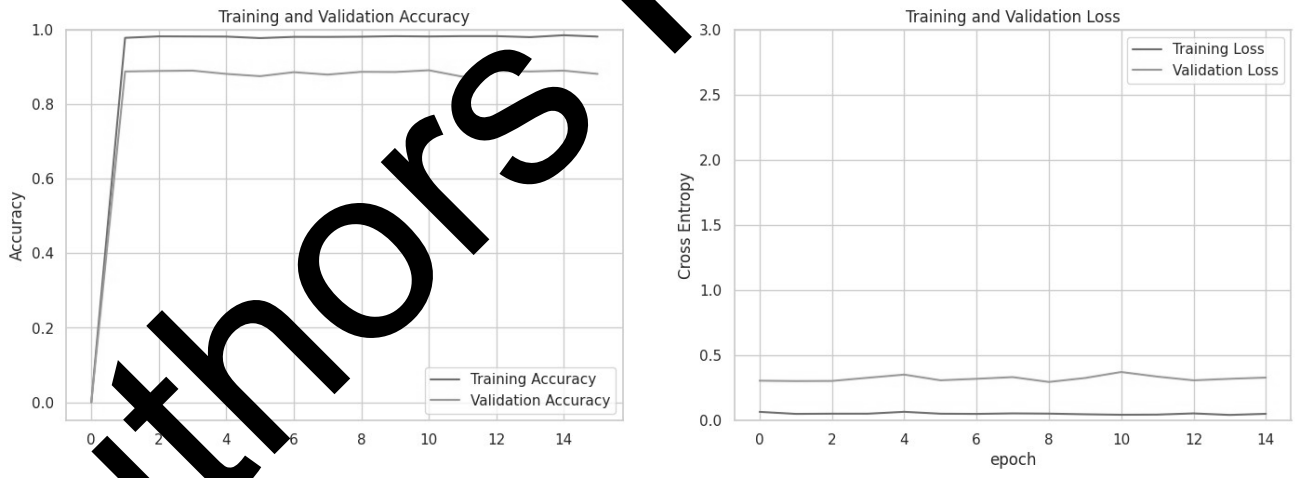


Fig. 5. Epoc-wise Training and Validation Accuracy Curve (Weighted ensemble)

We can identify patterns of overfitting, underfitting, or ideal model behaviour by examining the convergence or divergence of the two curves, which helps us make decisions about the optimization and refinement of the model. Notably, our ensemble model exhibited a remarkable AUC value of 0.98 in the classification of benign and melanoma skin cancer classes. The ROC curves for these classes are shown in Figure 6, which

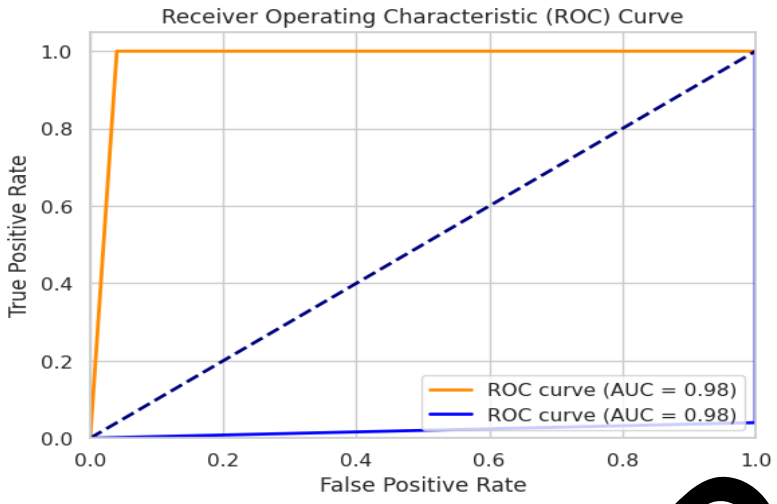


Fig. 6. AUC value of weighted ensemble mode

highlights the model's remarkable discriminatory ability. This outstanding demonstrati highlights how well our group method works to distinguish between benign and malignant skin lesions. The confusion ma x of the proposed weighted ensemble model is shown in Figure 7. Our suggested model performed remarkably ell in accurately and dependably classifying melanoma samples. The model successfully distinguished eac 250 melanoma samples with extreme precision, demonstrating its resilience in identifying cancerous lesions. lore er, the model demonstrated exceptional competence in classifying benign samples, correctly classifying 24 ou 250 c es. The model made some errors in classifying the beginning class.

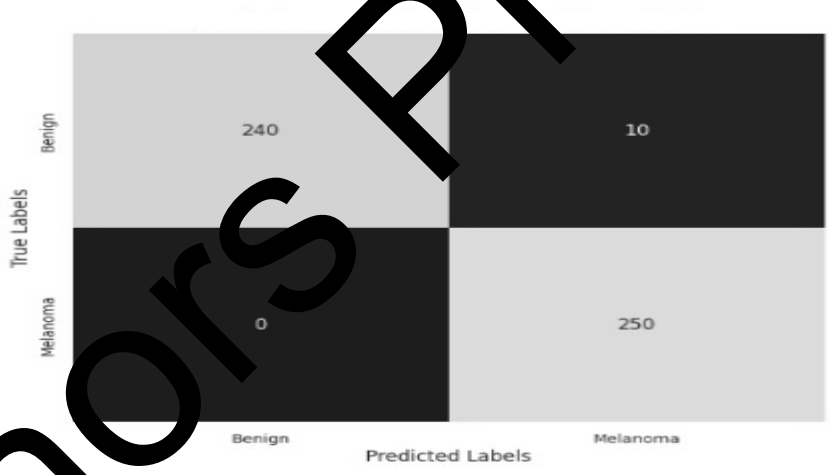


Fig. 7. Confusion matrix (Ensemble model)

VII. CONCLUSION

Early detecti and precise prediction of melanoma can greatly reduce mortality rates and increase survival rates. Recent impr ements kin disease recognition, specifically in early skin cancer detection, have demonstrated promising results with no l D p-Learning models. Our work contributes to this rapidly developing field by presenting a novel Transfer l ning model with an Ensemble Deep-Learning model designed for the classification of skin cancer. Even though a umber of models have shown remarkable accuracy rates, issues like unbalanced datasets, a lack of data, and low image quality still make optimal performance difficult to achieve. Our model seeks to tackle these issues by using augmentation techniques and overall model robustness enhancement. The discoveries of our study provide useful insights into the development of more effective skin disease recognition models. Our Transfer Learning model with an Ensemble-Deep Learning approach presents an optimistic solution to the challenges encountered in the field, aiming to enhance early detection rates and eventually save lives. Future research should concentrate on extended data augmentation techniques, analyze the benefit of hyperspectral images for richer data analysis, and work on improving model resilience and

adaptability to various scenarios. Resolving these problems can improve the detection of skin diseases, providing medical personnel with better resources for accurate and early diagnosis and improving patient outcomes.

REFERENCES

[1] F. Bray, M. Laversanne, B. Cao, C. Varghese, B. Mikkelsen, E. Weider-pass, and I. Soerjomataram, "Comparing cancer and cardiovascular dis- ease trends in 20 middle-or high-income countries 2000–19: A pointer to national trajectories towards achieving sustainable development goal target 3.4," *Cancer Treatment Reviews*, vol. 100, p. 102290, 2021.

[2] N. A. AlSadhan, S. A. Alamri, M. M. Ben Ismail, and O. Bchir, "Skin cancer recognition using unified deep convolutional neural networks," *Cancers*, vol. 16, no. 7, p. 1246, 2024.

[3] M. A. Rahman, E. Bazgir, S. S. Hossain, M. Maniruzzaman, *et al.*, "Skincancer classification using nasnet," *International Journal of Science andResearch Archive*, vol. 11, no. 1, pp. 775–785, 2024.

[4] B. Sarker, N. Sharif, M. Rahman, and A. Parvez, "Ai, iomt and blockchain in healthcare," *Journal of Trends in Computer Science and Smart Technology*, vol. 5, no. 1, pp. 30–50, 2023.

[5] A. Naeem, T. Anees, M. Khalil, K. Zahra, R. A. Naqvi, and S.-W.Lee, "Snc net: Skin cancer detection by integrating handcrafted and deep learning-based features using dermoscopy images," *Mathematics*, vol. 12, no. 7, p. 1030, 2024.

[6] K. Monica, J. Shreeharsha, P. Falkowski-Gilski, B. Falkowska-Gilska, M. Awasthy, and R. Phadke, "Melanoma skin cancer detection using mask-rcnn with modified gru model," *Frontiers in Physiology*, vol. 14, p. 1324042, 2024.

[7] M. T. Campos, F. A. Silva, J. R. Fernandes, S. G. Santos, F. D. Magalhães, M. J. Oliveira, and A. M. Pinto, "New mos2/tegafur- containing pharmaceutical formulations for selective led-based skin cancer photo-chemotherapy," *Pharmaceutics*, vol. 16, no. 3, p. 360, 2024.

[8] A. Naeem and T. Anees, "Dvfnet: A deep feature fusion-based model for the multiclassification of skin cancer utilizing dermoscopy images,"*Plos one*, vol. 19, no. 3, p. e0297667, 2024.

[9] N. A. AlSadhan, S. A. Alamri, M. M. Ben Ismail, and O. Bchir, "Skin cancer recognition using unified deep convolutional neural networks," *Cancers*, vol. 16, no. 7, p. 1246, 2024.

[10] Puttaswamy, B. S., and N. Thillaiarasu. "Fine DenseNet based human personality recognition using english hand writing of non-native speakers." Biomedical Signal Processing and Control 99 (2025): 106910.

[11] I. Kousis, I. Perikos, I. Hatzilygeroudis, and M. Virvou, "Deep learningmethods for accurate skin cancer recognition and mobile application," *Electronics*, vol. 11, no. 9, p. 1294, 2022.

[12] A. Rajesh, K. N. Rao, G. N. V. Sai, K. D. Kumar, and P. R. S. Karthik, "Skin cancer detection and intensity analysis using deep learning," in *2024 International Conference on Emerging Systems and Intelligent Computing (ESIC)*, pp. 376–381, IEEE, 2024.

[13] M. S. Sivakumar, L. M. Leo, T. Gurumekala, V. Sindhu, and A. S. Priyadharshini, "Deep learning in skin lesion analysis for malignant melanoma cancer identification," *Multimedia Tools and Applications*, vol. 83, no. 6, pp. 17833–17853, 2024.

[14] Ashwin Shenoy, M., and N. Thillaiarasu. "Enhancing temple surveillance through human activity recognition: A novel dataset and YOLOv4-ConvLSTM approach." Journal of Intelligent & Fuzzy Systems Preprint (2023): 1-16.

[15] J. V. Tembhurne, N. Hebbar, H. Y. Patil, and T. Diwan, "Skin cancer detection using ensemble of machine learning and deep learning tech- niques," *Multimedia Tools and Applications*, vol. 82, no. 18, pp. 27501–27524, 2023.

[16] A. Shahsavari, T. Khatibi, and S. Ranjbari, "Skin lesion detection using an ensemble of deep models: Slded," *Multimedia Tools and Applications*, vol. 82, no. 7, pp. 10575–10594, 2023.

[17] M. Shorfuzzaman, "An explainable stacked ensemble of deep learning models for improved melanoma skin cancer detection," *Multimedia Systems*, vol. 28, no. 4, pp. 1309–1323, 2022.

[18] K. M. Saravanan, Gnanagurusubbiah, R. R. R. Roy, S. Balu, *et al.*, "Enhancing skin lesion classification with advanced deep learning ensemble models: a path towards accurate medical diagnostics," *CurrentProblems in Cancer*, p. 101077, 2024.

[19] K. Ali, Z. A. Shaikh, A. A. Khan, and A. A. Laghari, "Multiclass skin cancer classification using efficientnets–a first step towards preventing skin cancer," *Neuroscience Informatics*, vol. 2, no. 4, p. 100034, 2022.

[20] T. M. Ali, K. Shaukat, W. A. Khan, I. A. Hameed, L. A. Almuqren, M. A. Raza, M. Aslam, and S. Luo, "An efficient deep learning-based skin cancer classifier for an imbalanced dataset," Diagnostics, vol. 12, no. 9, p. 2115, 2022.

[21] A. Ali, F. Yilmaz, and O. Kose, "Early detection of skin cancer using deep learning architectures: resnet-101 and inception-v3," in 2019 medical technologies congress (TIPTEKNO), pp. 1–4, IEEE, 2019.

[22] Ravi Prasad, M., and N. Thillaiarasu. "Multichannel EfficientNet B7 with attention mechanism using multimodal biometric-based authentication for ATM transaction." Multiagent and Grid Systems 20.2 (2024): 89-108.

[23] K. M. Hosny, M. A. Kassem, and M. M. Foaud, "Skin cancer classi- fication using deep learning and transfer learning," in 2018 9th Cairo international biomedical engineering conference (CIBEC), pp. 90–93, IEEE, 2018.