A Framework for Skin Cancer Classification using Ensemble Transfer Learning and Data Fusion

Tuba Mubarik, Shabib Aftab, Muhammad Anwaar Saeed
Department of Computer Science, Virtual University of Pakistan, Lahore, Pakistan
*Correspondence:mubariktuba.1205@gmail.com,shabib.aftab@gmail.com,anwaar@vu.edu.pk

Citation | Mubarik. T, Aftab. S, Saeed. M. A, "A Framework for Skin Cancer Classification using Ensemble Transfer Learning and Data Fusion", IJIST, Vol. 8 Issue. 5 pp 2113-2133, August 2026
DOI | https://doi.org/10.33411/IJIST/1991
Received | July 28, 2026 **Revised |** Aug 17, 2026 **Accepted |** Aug 18, 2026 **Published |** Aug 21, 2026.

Skin cancer is a serious and growing health problem worldwide. Its cases keep rising every year. Early diagnosis plays a key role in improving survival. Manual diagnosis through visual checks and dermoscopy depends on the doctor's experience. It can also give inconsistent results. Biopsy is the most reliable method. But it is invasive and time-consuming. These limits have pushed researchers toward deep learning-based systems. Most past studies use only one model. This lowers accuracy when skin lesions vary in colour, texture, and shape. To fix this, this study proposes an ensemble model combining EfficientNetB3 and Xception. The ISIC 2016 and ISIC 2017 datasets were combined for more diverse training data. Class imbalance was fixed using SMOTE. Both models were trained using transfer learning with ImageNet weights. They were trained under 3-fold cross-validation on a fixed test set. Mixup augmentation was also used during training. Predictions from both models were combined using equal-weighted averaging. The model was evaluated using accuracy, precision, recall, ROC-AUC, and confusion matrices. The ensemble reached a test accuracy of 95.22%, outperforming several existing methods. The ensemble also reached a sensitivity of 92.57%, a specificity of 97.88%, and a ROC-AUC of 0.9878. It improved over the standalone Xception model, which reached 95.01% accuracy, by 0.21 percentage points. It also improved over the standalone EfficientNetB3 model, which reached 94.90% accuracy, by 0.32 percentage points. These results confirm that this ensemble approach, with SMOTE-based balancing, offers a reliable solution for automated skin cancer classification.

Keywords: Skin Cancer Classification, Deep Learning, Ensemble Learning, EfficientNetB3, Xception.

Introduction:

Skin cancer is one of the most common types of cancer and has become a serious health problem around the world. The number of patients is increasing every year, and many cases are diagnosed at an advanced stage [1]. Skin cancer develops when skin cells start growing in an uncontrolled way. Long-term exposure to ultraviolet (UV) radiation is considered the main cause of this disease. Other factors, such as age, skin type, family history, and environmental conditions, also increase the risk of developing skin cancer. The disease mainly includes basal cell carcinoma (BCC), squamous cell carcinoma (SCC), and melanoma. Among these, melanoma is the most dangerous type [2]. It can spread quickly to other parts of the body. This happens if it is not detected early. Early diagnosis increases the chances of successful treatment and improves the survival rate of patients [3]. Because of this, researchers are focusing more on computer-based systems [4]. These systems can help doctors detect skin cancer quickly and accurately.

The diagnosis of skin cancer usually starts with a visual examination performed by a dermatologist. In many cases, dermoscopy is used to examine skin lesions in more detail. It helps doctors observe structures that cannot be seen with the naked eye [5]. Although dermoscopy improves the examination process, the final diagnosis still depends on the experience of the specialist. In many situations, benign and malignant skin lesions look very similar. They may have similar colours, shapes, and textures, which makes diagnosis difficult [6]. A biopsy is the most reliable method for confirming skin cancer. But it is invasive. . It also requires laboratory testing and additional clinical resources This process may also increase the diagnosis time [7]. Because of these challenges, researchers are developing automated systems that can support dermatologists by providing faster and more consistent diagnostic results [8][9].

In recent years, artificial intelligence has become an important part of medical image analysis. Among different AI techniques, deep learning has shown excellent performance in image classification. These models automatically learn useful features from medical images. This reduces the need for manual feature extraction [10]. Convolutional Neural Networks (CNNs) are successfully used for disease detection and classification [10][11]. Transfer learning has further improved this field. It allows pretrained models to adapt for new medical applications. This is very helpful when training data is limited. At the same time, ensemble learning has become a popular approach. It combines predictions from multiple models to improve final decisions [12].

Similar approaches are successful in other healthcare applications. For cardiovascular disease, models are used to predict risk based on patient data and imaging features [13]. For diabetic retinopathy, deep learning models are used to detect the disease early from retinal images [14]. These models have also been extended to grade the severity of diabetic retinopathy, helping doctors decide on the right stage of treatment [15]**.** These methods improve prediction performance and support clinical decision-making.

CNN and deep learning-based models are not limited to medical imaging alone. They are also used in other fields, such as software engineering. For example, data and decision-level fusion techniques have been applied to predict software defects before deployment [16]. This shows that ensemble and fusion-based deep learning strategies generalize well across very different domains, not just healthcare. This further supports the reliability of these techniques when applied to complex classification problems such as skin cancer detection. Ultimately, these techniques can be applied effectively to skin cancer classification.

Publicly available datasets have played an important role in the development of automated skin cancer classification systems [17][18]. Among them, the International Skin Imaging Collaboration (ISIC) datasets are the most widely used benchmark datasets for skin lesion analysis [19][20]. The ISIC 2016 dataset introduced one of the earliest international

challenges for automated skin lesion classification and has been used in many deep learning studies. The ISIC 2017 dataset later provided a larger collection of annotated dermoscopic images and became another standard benchmark for evaluating skin lesion classification models. These datasets contain images collected under different conditions and have helped researchers compare different deep learning methods using the same evaluation standards. As a result, they are widely used to develop and evaluate automated skin cancer detection systems.

Many researchers have proposed deep learning models for skin cancer classification, and several studies have reported promising results [21][8]. However, many existing methods still depend on a single deep learning model. The performance of a single model may decrease when skin lesions show large differences in colour, texture, size, and shape. Different deep learning models learn different image features, and each model has its own strengths and limitations. Ensemble learning provides an effective solution by combining the predictions of multiple models and reducing the weaknesses of individual networks [22]. Ensemble learning has shown good performance in different medical image tasks. But few studies have combined EfficientNetB3 [23] and Xception [24] together. Even fewer have tested this combination on the ISIC 2016 and ISIC 2017 datasets for binary skin cancer classification. This research gap motivated us to develop a new ensemble model that can provide better classification performance.

In this study, an ensemble transfer learning approach is proposed for binary skin cancer classification using the ISIC 2016 and ISIC 2017 datasets. First, all images are preprocessed to prepare a consistent dataset for training. The two datasets are then combined to increase the diversity of training samples. Class imbalance is handled using the Synthetic Minority Oversampling Technique (SMOTE) [25], and data augmentation is applied to improve model learning [26]. Two pretrained deep learning models, EfficientNetB3 and Xception, are trained separately using transfer learning. Their predictions are then combined through an ensemble strategy to produce the final classification result. The proposed model is evaluated using different performance measures. These include accuracy, precision, recall, ROC-AUC, confusion matrices, and learning curves. The experimental results show that the proposed ensemble model achieved an accuracy of 95.22%. This shows that the proposed approach can provide reliable performance for automated skin cancer classification. Therefore, the objective of this study is to develop a reliable ensemble transfer learning framework for binary skin cancer classification. The proposed framework combines two public datasets to improve the classification performance. Ensemble learning is also used to combine EfficientNetB3 and Xception for better performance.

The main contribution of this work is an ensemble transfer learning model. It combines EfficientNetB3 and Xception for binary skin cancer classification. It uses the ISIC 2016 and ISIC 2017 datasets. The study also investigates the effect of combining two public datasets, applying SMOTE for class balancing, and using data augmentation to improve the learning process. A comprehensive performance evaluation is carried out using several standard metrics to demonstrate the effectiveness of the proposed approach.

Research Novelty:

Unlike previous studies that use just one model or one dataset this study uses one framework. This framework combines dataset-level fusion, SMOTE-based class balancing, data augmentation and an ensemble of EfficientNetB3 and Xception. This mix helps make training more robust and makes the binary skin cancer classification more reliable.

The proposed framework is also different from other ensemble methods used for skin cancer classification. [12] and Halder et al. [27] combined three CNNs in their ensembles. [28] combined three CNNs with patient metadata. [29] paired a CNN with an LSTM-ELM ensemble. These methods use three or more models, or need extra metadata that is not always available. This increase training cost and the risk of overfitting on datasets of moderate size.

The proposed framework instead uses only two CNNs, EfficientNetB3 and Xception, combined with simple equal-weighted averaging. It does not need patient metadata. It also combines two public datasets, ISIC 2016 and ISIC 2017, before training, instead of only combining predictions. [30] and [31] used classical machine learning methods such as SVM and Random Forest inside their ensembles. The proposed framework keeps the ensemble fully within deep transfer learning models, which keeps training simpler. To our knowledge, this combination of dataset fusion, SMOTE, Mixup, and a two-model EfficientNetB3 and Xception ensemble has not been tested together before.

Related work:

Deep learning has become one of the most widely used techniques for skin cancer classification. Early studies mainly focused on Convolutional Neural Networks (CNNs) because of their ability to learn image features automatically. These models reduced the need for manual feature extraction and improved the performance of skin lesion classification. As research progressed, several CNN architectures were introduced to improve classification accuracy. [21] proposed a two-stage residual network. It segmented and classified melanoma on the ISIC 2016 dataset. [8] trained a CNN on dermoscopic images and compared its performance directly against 145 dermatologists. [23] introduced Efficient Net, which many later studies adapted for skin lesion classification. These studies showed that deep learning can support dermatologists by providing fast and reliable predictions. However, the performance of these models often depended on the quality of the training data and the architecture selected for classification.

Transfer learning further improved the performance of deep learning models by using knowledge learned from large image datasets. Instead of training a model from the beginning, pretrained models were fine-tuned using medical images. This approach reduced training time and improved classification performance, especially when only a limited number of medical images were available. Transfer learning has also been successfully used in other medical image analysis applications. [13] used a data and machine learning fusion architecture for cardiovascular disease prediction. [14] proposed a transfer learning and data fusion framework for diabetic retinopathy detection. [15] extended this idea into a severity classification framework using ensemble transfer learning. Studies based on transfer learning, data fusion, and ensemble learning demonstrated that combining different learning strategies can improve prediction accuracy and produce more reliable results. These findings encouraged researchers to apply similar techniques to skin cancer classification.

In recent years, ensemble learning has received increasing attention because it combines the predictions of multiple deep learning models instead of relying on a single network. Different models learn different image features, and their combination often produces better classification performance than an individual model.

Several studies have applied this idea to skin cancer classification. [12] fused three CNNs trained on cropped images at different scales and reached 86.2% accuracy on ISIC 2018. [32] combined four deep networks for melanoma and seborrheic keratosis detection on ISIC 2017, reaching a mean AUC of 0.958. [30] used a Max Voting ensemble of Random Forest, SVM, and a neural network with genetic algorithm-based feature selection, reaching 94.70% on HAM10000 and ISIC 2018. [28] fused ResNet50, Xception, and EfficientNetB0 with patient metadata and reached 93.2% on ISIC 2018. [27] combined Xception, InceptionResNetV2, and MobileNetV2 using fuzzy logic and reached 95.14% on HAM10000. [29] paired a Squeeze-Excitation Dense Net with an optimized LSTM-ELM ensemble and reached about 95.1% on HAM10000 and ISIC datasets. [31] combined a CNN with an SVM and reached 92% on HAM10000.

More recent work has extended this idea further. It combines Efficient Net backbones with attention-based Vision Transformer boosters [33]. It also uses metaheuristic optimization

strategies [34]. This further improves ensemble performance in skin lesion classification. Ensemble learning has also been combined with transfer learning to improve medical image classification, showing strong results in applications such as diabetic retinopathy severity classification. Similar fusion-based ensemble strategies have also proven effective outside healthcare, such as in software defect prediction, showing that these techniques generalize well across different types of classification problems. Most of these ensemble models combine three or more networks. This raises training cost and computational complexity.

The International Skin Imaging Collaboration (ISIC) datasets have become the standard benchmark for evaluating automated skin cancer classification models. The ISIC 2016 and ISIC 2017 datasets have been used by many researchers to compare different deep learning approaches under the same evaluation settings. [35] combined EfficientNetB3 with an XGBoost classifier for clinical metadata on the PAD-UFES-20 dataset. Recent frameworks have also combined Efficient Net with wavelet-based preprocessing and robust multiclass architectures to achieve further gains in classification accuracy on skin lesion datasets [36][37].

More recent studies have investigated attention mechanisms and feature fusion to improve classification accuracy. [38] proposed an attention-based mechanism to combine images and clinical metadata. [39] used dense convolutional networks with fused metadata for skin tumor classification. [40] showed that integrating clinical data with imaging improves diagnostic accuracy. [41] compared transformers and CNNs on small skin lesion datasets. Although these methods achieved promising results, many of them were evaluated using a single dataset or a single deep learning architecture. As a result, there is still room to improve model reliability by combining complementary deep learning models.

Based on the existing literature, it can be observed that deep learning, transfer learning, and ensemble learning have significantly improved automated skin cancer classification. However, several challenges still remain. [42] used Siamese neural networks for skin cancer classification and new class detection but relied on a single dataset. [43] proposed SkinLesNet, a custom multi-layer CNN, which was also evaluated on only one dataset. [44] compared several machine learning and deep learning models and found that even the best model depended heavily on the choice of dataset. Ensemble learning based on CNNs has shown strong results, but the field has also moved toward newer methods in recent years. It is useful to look at these methods and see where they stand compared to the proposed approach. Vision Transformers are one such method and are becoming popular for skin lesion analysis. [45] used a Vision Transformer with automatic segmentation on the HAM10000 dataset and reached over 96% accuracy. [40] added anatomical site information to a Vision Transformer for skin cancer classification. [46] combined a Vision Transformer with a CNN to capture both global and local lesion features. Along with better accuracy, doctors also need to trust a model's decision, which is why explainable AI is getting more attention. Grad-CAM is a common method used to show which part of an image influenced a model's decision [47]**.** [48] reviewed many studies that used Grad-CAM in medical imaging, including skin lesion classification. Foundation models are another newer direction. These models are trained on very large sets of medical images and can be adapted to different tasks with less labelled data [49]**.** These newer methods often need larger datasets or more computing power than a typical research setting provides. This is why the present study still uses a CNN-based ensemble, which needs less data and less computing power. It remains a practical choice for datasets of this size, such as ISIC 2016 and ISIC 2017. The diversity of skin lesion images and class imbalance continue to affect classification performance. Also, only a few studies have combined EfficientNetB3 and Xception together. Even fewer studies test this combination on ISIC 2016 and ISIC 2017. These are two of the most widely used skin lesion datasets [19-20]. This shows a clear gap that this study aims to fill.

Table 1. lists the main gaps found in past studies and shows how the proposed framework responds to each one.

Gap Found	Related Studies	How this study responds
Uses only one dataset	[42][43][44]	Combines ISIC 2016 and ISIC 2017 datasets before training
Ensembles use three or more CNNs	[28][27]	Uses only two CNNs with equal-weighted averaging
Needs patient metadata	[28][35][38][40]	Uses only dermoscopic images
Few studies test EfficientNetB3 and Xception together	[23][24]	Tests this pair directly on ISIC 2016 and ISIC 2017
Class imbalance handled differently across studies	[30][31]	Uses SMOTE, Mixup and class weight together

To address these limitations, this study proposes an ensemble transfer learning approach that combines EfficientNetB3 and Xception for binary skin cancer classification. The proposed method also includes image preprocessing, dataset combination, SMOTE-based class balancing, and data augmentation to improve the learning process and achieve better classification performance.

Material and Methods:

Datasets:

This study uses two public skin lesion datasets: ISIC 2016 and ISIC 2017. Both datasets contain dermoscopic images acquired under clinical conditions using dermoscopy devices. This makes the two datasets easy to combine without style mismatch. ISIC 2016 contains 900 images in total. Each image is labelled as benign (0) or malignant (1). This dataset has 727 benign images and 173 malignant images. The benign class is much larger than the malignant class in this dataset.

ISIC 2017 contains 2,000 images in total. It does not give a single benign or malignant label directly. Instead, it gives three separate labels for each image: melanoma, seborrheic keratosis, and nevus. These labels needed to be converted into a binary format for this study. Images with a melanoma label were marked as malignant (1). All other images, including seborrheic keratosis and nevus cases, were marked as benign (0). Under this rule, ISIC 2017 gives 1,626 benign images and 374 malignant images. Combining both datasets increase the total amount of training data available. It also adds more variety in lesion appearance, lighting conditions, and image quality, which helps the model learn more general features.

Site Map: This is a computational study based on two publicly available image datasets, ISIC 2016 and ISIC 2017. No physical field site or clinical location was involved in data collection for this work.

Image Preprocessing:

Images from both datasets are read and resized to 224×224 pixels. This matches the input size needed by the models. Images are also converted from BGR to RGB format. Both EfficientNetB3 and Xception require RGB input. Each image is matched to its label using its filename. Any image with no matching label is removed. Images that fail to load are also removed. Fixed random seeds are set for Python, NumPy, and TensorFlow. This ensures the same results across different runs. It also keeps the full pipeline reproducible.

Dataset Fusion, Balancing and Splitting:

After preprocessing, ISIC 2016 and ISIC 2017 are merged. This creates one dataset of 2,900 images. It contains 2,353 benign and 547 malignant images. This shows a clear class imbalance. Benign images outnumber malignant ones by more than four to one. To fix this, SMOTE is applied to the combined dataset. SMOTE creates new malignant samples through interpolation. It does not simply duplicate existing images. This process continues until both

classes reach 2,353 samples. The final balanced dataset contains 4,706 images. This balancing step happens before any data splitting. This ensures both classes are equal from the start.

The balanced dataset is shuffled first. This removes any ordering bias in the data. It is then split into two parts using an 80-20 ratio. Stratified sampling is used for this split. This keeps the class ratio equal in both parts. The training set contains 3,764 images, making up 80% of the data. The fixed test set contains 942 images, making up the remaining 20%. The training set is further divided using 3-fold stratified cross-validation. In each fold, two parts are used for training and the remaining part is used for validation. This means a separate fixed validation split is not needed, since every fold performs its own validation internally. The same three folds are used for both EfficientNetB3 and Xception. This keeps the comparison between the two models unbiased. The test set stays separate from all folds. It is used only to evaluate the trained models. The same test set is reused for every fold. It is also reused for both models. This makes all reported results directly comparable.

Data Augmentation and Class Weighting:

Since medical image datasets are often limited in size, extra augmentation helps the model learn better. It is applied to the training data inside each fold only. First, the training images in each fold are duplicated. This increases the number of samples available for training. During training, a Mixup-based technique is used instead [50]. This replaces traditional augmentation methods. Mixup blends pairs of training images together. It also blends their corresponding labels at the same time. The blending ratio is drawn from a Beta distribution ($\alpha = 0.2$). This creates new training examples. These examples sit between two real images. It helps the model learn smoother decision boundaries. It also reduces the risk of overfitting.

To handle the remaining class imbalance, a class weighting strategy is used. Class weights are calculated separately for each fold. They follow a balanced weighting scheme. These weights are applied throughout training, alongside Mixup. This gives the malignant class fair importance during learning. This step matters even after SMOTE balancing. Without it, the model could still lean toward the majority class during fold-level training.

Proposed Framework:

The proposed framework is built on two layers. The first layer is the Training Layer. It handles data loading and preprocessing. It also handles class balancing and model training. The second layer is the Testing Layer. It handles prediction and final classification. Both layers are connected through one interface. This interface is the trained model weights. The Training Layer produces these weights as output. The Testing Layer then uses them as input. Figure 1 illustrates the complete architecture of the proposed framework.

As shown in Figure 1, each layer performs a distinct set of operations.

In the Training Layer, ISIC 2016 and ISIC 2017 are combined first. All images are preprocessed before training. This includes resizing, RGB conversion, and label matching. SMOTE is then used to balance the classes. The data is split into a training set and a fixed test set, using an 80-20 ratio. The 80 percent is for training and the remaining 20 percent is for testing. The training set is divided into three folds. For each fold, EfficientNetB3 and Xception are trained separately. Both use transfer learning with ImageNet weights. Both also use Mixup augmentation and class weighting. Training uses the Adam optimizer. Three callbacks are used during training. These are Model Checkpoint, Early Stopping, and Reduce LROnPlateau.

In the Testing Layer, all trained fold models make predictions. These predictions are made on the fixed test set. This is the same 20 percent test set that was set aside earlier. Predictions from the three folds of each model are averaged first. This gives one stable prediction per architecture. The two averaged predictions are then combined equally. This produces the final benign or malignant decision for each image.

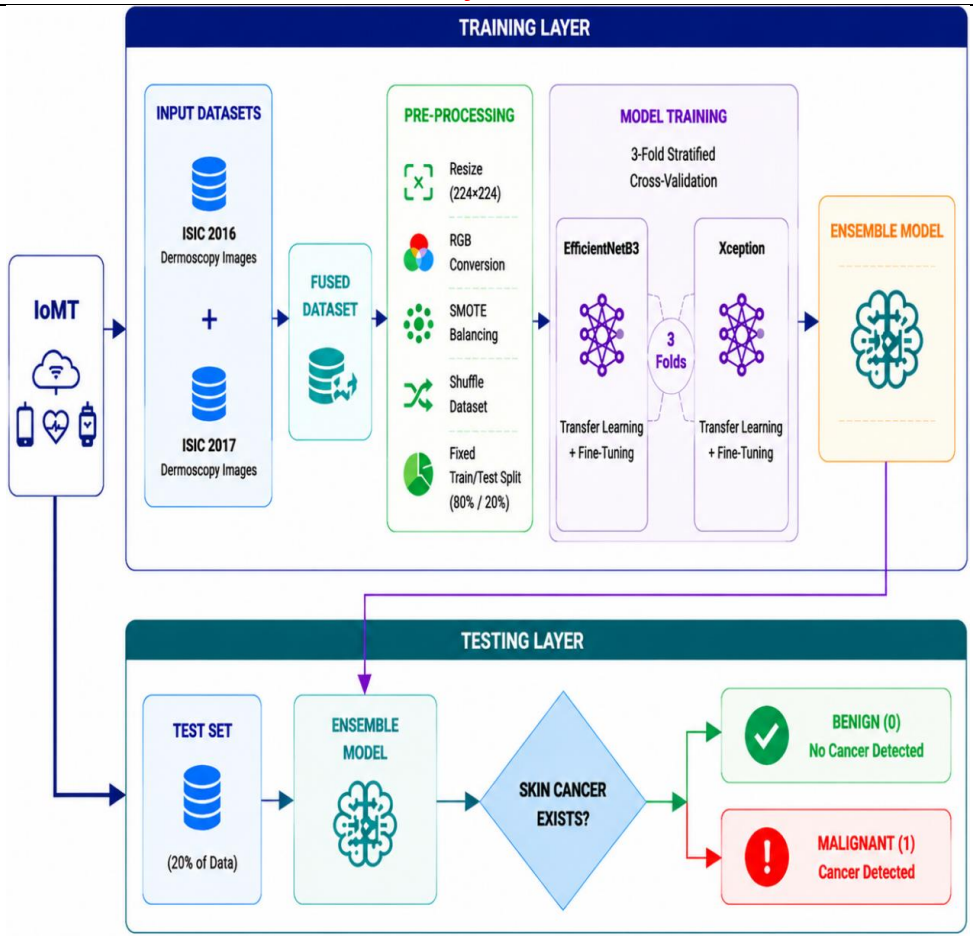

Figure 1. Proposed framework for Binary Classification

Model Architecture and Fine-Tuning:

Two pretrained convolutional neural network architectures are used for binary skin cancer classification: EfficientNetB3 and Xception. Both models are initialized with weights pretrained on ImageNet. A single-phase partial fine-tuning strategy is used for both models. In this strategy, the first 60% of layers are frozen from the start of training. The remaining 40% of layers stay trainable throughout training. This lets both models keep their general ImageNet features. At the same time, it allows them to adapt to skin lesion patterns.

For EfficientNetB3, a partial fine-tuning strategy is applied. The first 60% of layers are kept frozen, preserving general features learned from ImageNet. The remaining 40% of layers are fine-tuned on the skin lesion data, allowing the model to adjust to the new task. A custom classification head is then added on top of this base. It starts with a global average pooling layer, followed by a batch normalization layer. Next, a dense layer with 128 units and ReLU activation is applied, followed by a dropout layer with a rate of 0.3. A second batch normalization layer comes after this, followed by another dense layer with 64 units and ReLU activation, and a second dropout layer with a rate of 0.2. Finally, the head ends with a single dense layer using sigmoid activation, which produces the final binary output. EfficientNetB3 uses the resized pixel values directly, without any extra preprocessing.

For Xception, the same fine-tuning strategy is applied. As with EfficientNetB3, the first 60% of layers are kept frozen to preserve general ImageNet features, while the remaining 40% of layers are fine-tuned on the skin lesion data. The same custom classification head is added on top of this base. It starts with a global average pooling layer, followed by a batch normalization layer, then a dense layer with 128 units and ReLU activation, and a dropout layer with a rate of 0.3. This is followed by a second batch normalization layer, another dense

layer with 64 units and ReLU activation, and a second dropout layer with a rate of 0.2. The head ends in the same way, with a single dense layer using sigmoid activation for the final binary output. Unlike EfficientNetB3, however, Xception images are additionally passed through its own preprocessing function, matching the normalization used during its original ImageNet training.

Training Configuration:

Both models are trained using the Adam optimizer. The learning rate is set to 0.00005. Binary cross-entropy is used as the loss function. Training uses a batch size of 32. It runs for a maximum of 50 epochs. Three callbacks are used to keep training stable. ModelCheckpoint saves the weights with the best validation accuracy. EarlyStopping monitors the validation loss during training. It stops training if there is no improvement for 10 epochs. It also restores the best weights found so far. ReduceLROnPlateau lowers the learning rate when needed. It reduces the rate by half if validation loss stalls for 4 epochs. The learning rate can drop as low as 1×10^{-6}. Class weights are applied throughout training. This works alongside Mixup to handle any remaining imbalance.

Ensemble Strategy:

After training, each fold generates predictions on the test set. This includes all three EfficientNetB3 folds. It also includes all three Xception folds. For each architecture, the three-fold predictions are averaged first. This produces one stable prediction per architecture. The two averaged predictions are then combined equally. Neither architecture is given more weight than the other. A classification threshold of 0.5 is applied to this final score. This produces the final benign or malignant label for each test image.

Equal-weighted averaging was chosen instead of weighted averaging or stacking for three reasons. First, Xception and EfficientNetB3 reached close accuracy on their own, 95.01% and 94.90%. This small gap gives little reason to give one model more weight than the other. Second, equal-weighted averaging does not need extra parameters or an extra training step. Weighted averaging needs weights tuned on a validation set. Stacking needs a separate model trained on out-of-fold predictions. Adding this extra step raises the risk of overfitting on a dataset of this size. Third, equal-weighted averaging is simple to reproduce and explain, which matters for a tool meant to support clinical decisions.

Evaluation Metrics:

The ensemble model is evaluated using accuracy, misclassification rate, specificity, and sensitivity. Positive and negative predictive values are also calculated, along with false positive and false negative ratios. Positive and negative likelihood ratios are included as well. A confusion matrix is also generated for the results. It shows correct and incorrect predictions for each class. Training and validation accuracy curves are also plotted. Loss curves are plotted in the same way. These curves are averaged across all three folds. They help identify signs of overfitting or underfitting during training. ROC-AUC is also calculated to evaluate the overall separability between benign and malignant classes.

Computational Complexity:

Training was done on one Tesla P100-PCIE-16GB GPU. Training used about 15.5 GB of the GPU's 16GB memory. EfficientNetB3 took about 39 seconds per epoch on average. Xception took about 46 seconds per epoch. Across all three folds, EfficientNetB3 trained for 139 epochs in total, about 93 minutes. Xception trained for 101 epochs, about 78 minutes. Together, training took about 2.85 hours on one GPU. The EfficientNetB3 backbone weights are about 42 MB. The Xception backbone weights are about 80 MB. EfficientNetB3 has about 12 million parameters, and Xception has about 22 million**.** Testing all six-fold models on the fixed 942 image test set took about 68 seconds in total close to 14 images per second. This shows the ensemble can run on a single mid-range GPU without

special hardware. Direct comparison with other papers was not possible, since most of them do not report their own training time or hardware.

Results and Discussion:

There are a total of 900 images in the ISIC 2016 dataset. There are 2,000 images in the ISIC 2017 dataset. The fusion of these two datasets results in a total of 2,900 images. This gives 2,353 benign lesions and 547 malignant lesions. SMOTE was applied to this combined dataset. This step fixed the class imbalance. It resulted in a balanced dataset of 4,706 images. This dataset has 2,353 benign and 2,353 malignant images. The dataset was split into a training set and a fixed test set. An 80-20 ratio was used for this split. This makes a total of 3,764 images in the training set. The fixed test set contains 942 images. The training set was further divided using 3-fold stratified cross-validation. Each fold used two parts for training. The remaining part was used for validation. Before training, the images in each fold were duplicated. Mixup-based augmentation was then applied during training. Both transfer learning models, EfficientNetB3 and Xception, were trained separately. Each model was trained across all three folds. This was done to predict the presence or absence of skin cancer. To evaluate the performance, the following measures were used:

$$\text{Misclassification Rate} = (RA_1/TR_0 + RA_0/TR_1) / (TR_0 + TR_1) \quad (1)$$

$$\text{Accuracy} = (RA_0/TR_0 + RA_1/TR_1) / (TR_0 + TR_1) \quad (2)$$

$$\text{Positive Prediction Value} = (RA_1/TR_1) / (RA_1/TR_1 + RA_1/TR_0) \quad (3)$$

$$\text{Negative Prediction Value} = (RA_0/TR_0) / (RA_0/TR_0 + RA_0/TR_1) \quad (4)$$

$$\text{Specificity} = (RA_0/TR_0) / (RA_0/TR_0 + RA_1/TR_0) \quad (5)$$

$$\text{Sensitivity} = (RA_1/TR_1) / (RA_1/TR_1 + RA_0/TR_1) \quad (6)$$

$$\text{False Positive Ratio} = 1 - \text{Specificity} \quad (7)$$

$$\text{False Negative Ratio} = 1 - \text{Sensitivity} \quad (8)$$

$$\text{Likelihood Ratio Positive} = \text{Sensitivity} / (1 - \text{Specificity}) \quad (9)$$

$$\text{Likelihood Ratio Negative} = (1 - \text{Sensitivity}) / \text{Specificity} \quad (10)$$

In these equations, TR_0 represents the target result of the benign class. TR_1 represents the target result of the malignant class. RA_0 represents the result achieved as benign. RA_1 represents the result achieved as malignant. The combined term RA_j/TR_i does not represent a mathematical division. Instead, it denotes the corresponding cell of the confusion matrix. It represents the number of samples whose target class is i and whose achieved result is j. For example, RA_0/TR_0 represents the number of true negatives. RA_1/TR_1 represents the number of true positives. RA_1/TR_0 represents the number of false positives. RA_0/TR_1 represents the number of false negatives.

With a total of 3,764 training images, Xception correctly predicted all 1,882 images as benign and all 1,882 images as malignant, at the fold-averaged epoch where the best validation weights were saved. Table 2 shows the training confusion matrix for Xception. Xception achieved a training accuracy of 100.00%.

Table 2. Training Confusion Matrix for Xception

N=3764 (No. of Samples)	Result Achieved RA_0 (Negative-0)	Result Achieved RA_1 (Positive-1)
INPUT: Target Result TR_0 (Benign-0)	1882	0
INPUT: Target Result TR_1 (Malignant-1)	0	1882

EfficientNetB3 was trained using the same 3,764 images. At its best-weight epoch, it correctly predicted 1,882 images as benign and 1,860 images as malignant. Table 3 shows the training confusion matrix for EfficientNetB3. EfficientNetB3 achieved a training accuracy of 99.40%.

Table 3. Training Confusion Matrix for EfficientNetB3

N=3764 (No. of Samples)	Result Achieved RA_0 (Negative-0)	Result Achieved RA_1 (Positive-1)
INPUT: Target Result TR_0 (Benign-0)	1882	0
INPUT: Target Result TR_1 (Malignant-1)	22	1860

After training, both models were evaluated on the fixed test set of 942 images. Xception correctly predicted 457 benign and 438 malignant images. Table 4 presents the testing confusion matrix for Xception. Xception achieved a testing accuracy of 95.01%.

Table 4. Testing Confusion Matrix for Xception

N=942 (No. of Samples)	Result Achieved RA_0 (Negative-0)	Result Achieved RA_1 (Positive-1)
INPUT: Target Result TR_0 (Benign-0)	457	14
INPUT: Target Result TR_1 (Malignant-1)	33	438

EfficientNetB3 correctly predicted 459 benign and 435 malignant images on the same test set. Table 5 presents the testing confusion matrix for EfficientNetB3. EfficientNetB3 achieved a testing accuracy of 94.90%.

Table 5. Testing Confusion Matrix for EfficientNetB3

N=942 (No. of Samples)	Result Achieved RA_0 (Negative-0)	Result Achieved RA_1 (Positive-1)
INPUT: Target Result TR_0 (Benign-0)	459	12
INPUT: Target Result TR_1 (Malignant-1)	36	435

When the predictions of Xception and EfficientNetB3 were combined into the final ensemble, the model correctly predicted all 1,882 benign and all 1,882 malignant images on the training set. Table 6 shows this training confusion matrix. The ensemble achieved a training accuracy of 100.00%.

Table 6. Training Confusion Matrix for Ensemble

N=3764 (No. of Samples)	Result Achieved RA_0 (Negative-0)	Result Achieved RA_1 (Positive-1)
INPUT: Target Result TR_0 (Benign-0)	1882	0
INPUT: Target Result TR_1 (Malignant-1)	0	1882

On the fixed test set, the ensemble correctly predicted 461 benign and 436 malignant images. Table 7 shows this testing confusion matrix. The ensemble achieved a testing accuracy of 95.22%

Table 7. Testing Confusion Matrix for the Ensemble

N=942 (No. of Samples)	Result Achieved RA_0 (Negative-0)	Result Achieved RA_1 (Positive-1)
INPUT: Target Result TR_0 (Benign-0)	461	10
INPUT: Target Result TR_1 (Malignant-1)	35	436

Table 8 summarizes the performance of the individual models and the proposed ensemble on both the training and testing datasets.

Xception and EfficientNetB3 achieved very similar results on the test set. Xception obtained a testing accuracy of 95.01%, while EfficientNetB3 achieved 94.90%. A more meaningful difference can be seen in the gap between training and testing accuracy. Xception reached a training accuracy of 100.00%, which decreased to 95.01% on the test set. This gives a gap of about 5.0 percentage points. EfficientNetB3 achieved a training accuracy of 99.40% and a testing accuracy of 94.90%, resulting in a slightly smaller gap of about 4.5 percentage points. This suggests that EfficientNetB3 generalizes slightly better to unseen images. The ensemble model achieved the highest testing accuracy of 95.22%. Although the improvement

is small, it indicates that combining the predictions of both models helps reduce individual prediction errors. As a result, the ensemble provides the best overall performance on the test set.

The results show that both transfer learning models learned useful features from the fused datasets. The two models learn different image features. This makes them suitable for ensemble learning. The ensemble combines their predictions to reduce individual errors, making the overall classification results more stable.

Figures 2 and 3 show the training and validation accuracy and loss curves of the Xception model. The training accuracy increased steadily and reached 100.00%, while the validation accuracy also remained high throughout the training process. At the same time, both training and validation loss decreased gradually and stabilized in the final epochs. The small gap between the training and validation curves indicates that the model learned effectively without severe overfitting.

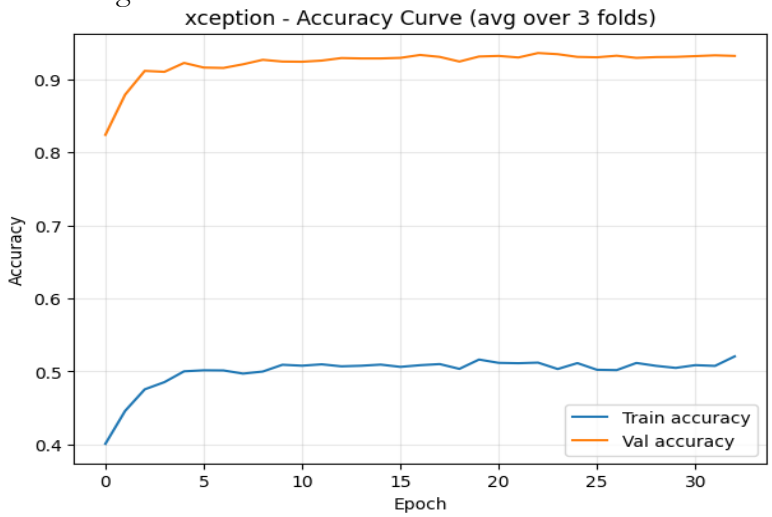

Figure 2. Training and validation accuracy for Xception

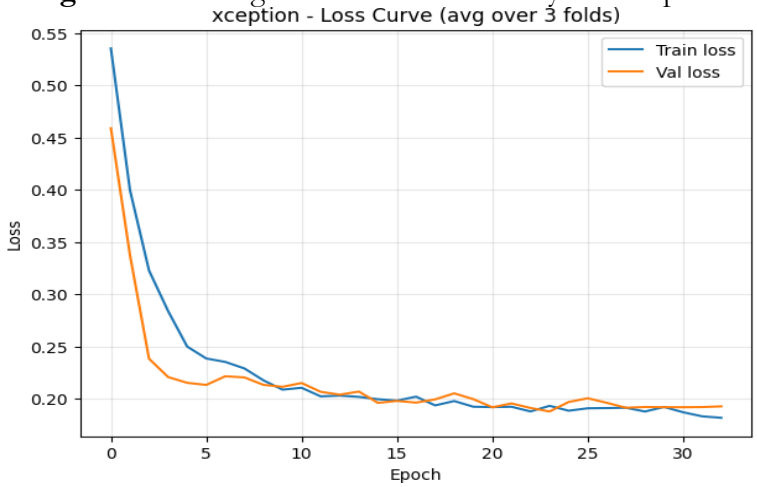

Figure 3. Training and validation loss for Xception

Table 8. Performance Evaluation of Individual and Ensemble Model

Model	Phase	Accuracy	Misclassification Rate	Specificity	Sensitivity	Positive Prediction Value	Negative Prediction Value	False Positive Ratio	False Negative Ratio	Likelihood Ratio Positive	Likelihood Ratio Negative
Xception	Training	1.0000	0.0000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	∞	0.0000
	Testing	0.9501	0.0499	0.9703	0.9299	0.9690	0.9327	0.0297	0.0701	31.2857	0.0722
Efficient NetB3	Training	0.9940	0.0060	1.0000	0.9874	1.0000	0.9886	0.0000	0.0126	∞	0.0126
	Testing	0.9490	0.0510	0.9745	0.9236	0.9732	0.9273	0.0255	0.0764	36.2500	0.0784
Ensemble	Training	1.0000	0.0000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	∞	0.0000
	Testing	0.9522	0.0478	0.9788	0.9257	0.9776	0.9294	0.0212	0.0743	43.6000	0.0759

Figures 4 and 5 present the training and validation accuracy and loss curves of the EfficientNetB3 model. The model showed a stable learning process, with both accuracy and loss curves converging smoothly during training. The validation curves remained close to the training curves, indicating good generalization on unseen data. Compared with Xception, EfficientNetB3 achieved a slightly smaller gap between training and testing accuracy, suggesting better generalization.

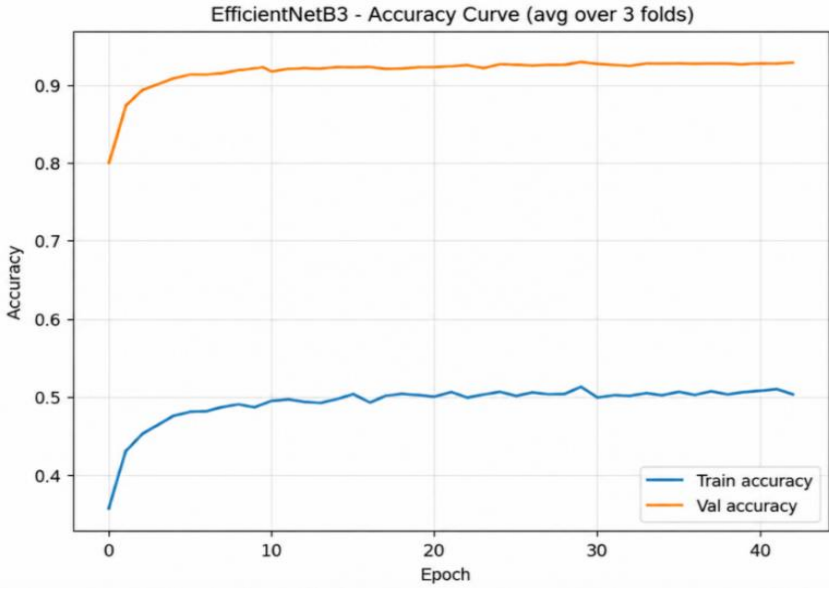

Figure 4. Training and validation accuracy for EfficientNetB3

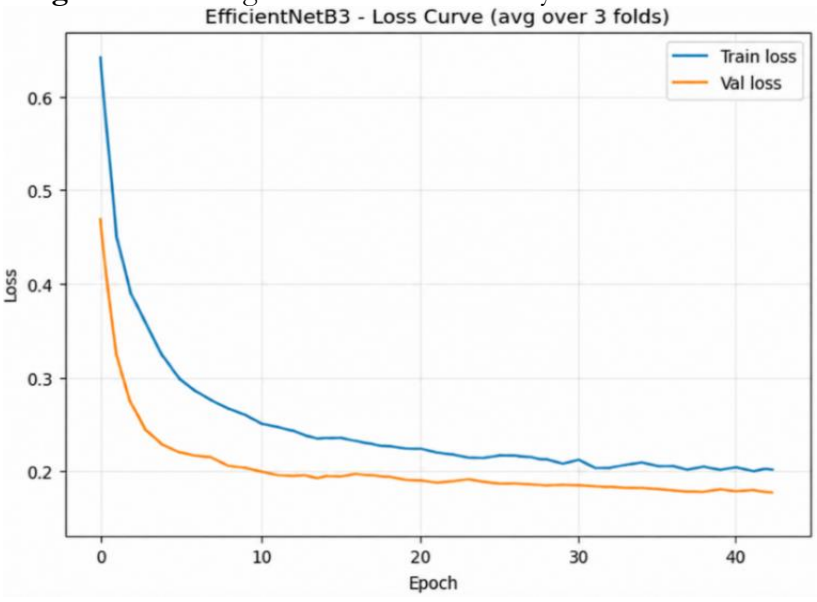

Figure 5. Training and validation loss for EfficientNetB3

Figures 6, 7, 8 present the ROC curves of Xception, EfficientNetB3, and the proposed ensemble model on the test set. All three models achieved high AUC values, indicating strong ability to distinguish between benign and malignant skin lesions. The ROC curves remain close to the upper-left corner, showing high sensitivity and specificity across different decision thresholds. Among the three approaches, the ensemble model achieved the highest AUC, demonstrating the best overall classification performance. This improvement indicates that combining the predictions of Xception and EfficientNetB3 enhances the discrimination capability of the model and produces more reliable predictions.

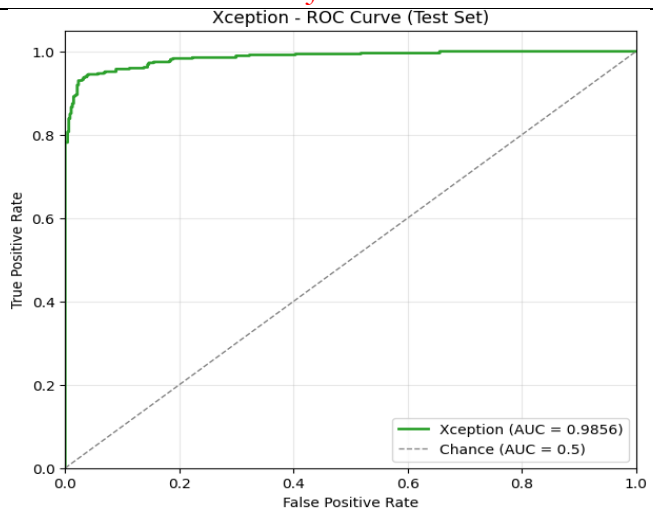

Figure 6. ROC curve graph for Xception Model

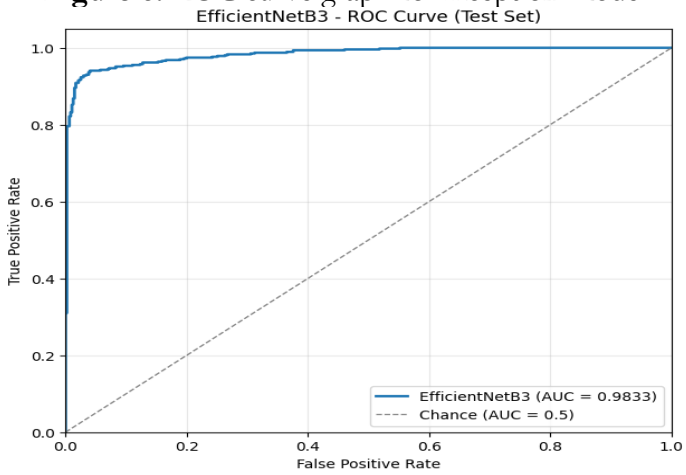

Figure 7. ROC curve graph for EfficientNetB3 Model

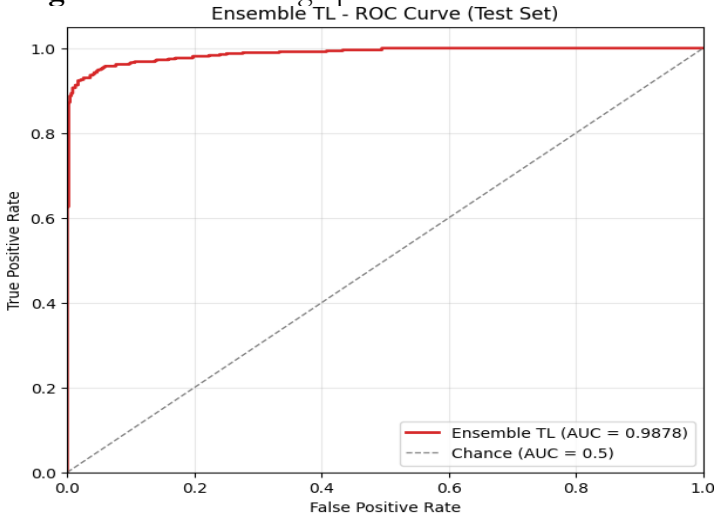

Figure 8. ROC curve graph for Ensemble

The high ROC-AUC values indicate that the proposed framework can effectively separate benign and malignant skin lesions over different decision thresholds. This shows that the model is not only accurate at a single threshold. It also stays stable under different classification conditions.

These results also matter in real clinical use. The ensemble had a sensitivity of 92.57%, meaning it correctly identified most malignant cases. A missed malignant case, called a false

negative, is the most dangerous error, since it delays treatment for the patient. Out of 471 malignant test images, the ensemble missed 35, giving a false negative rate of 7.43%, compared to 7.64% for EfficientNetB3 and 7.01% for Xception. A false positive is generally less critical, since it usually leads to additional clinical examination rather than delayed treatment. The ensemble produced only 10 false positives out of 471 benign images, giving a false positive rate of 2.12%, the lowest among all three models. This low false positive rate, combined with a high precision of 97.76%, shows that when the ensemble predicts malignant, it is very likely to be correct. These results also raise questions about real use in clinics. The ensemble can process the full 942 image test set in about 68 seconds on one GPU. This is fast enough for batch screening. A single image would return a result in a very short time once the models are loaded. This suggests the framework could support a triage role, flagging lesions for priority review, rather than replacing a dermatologist. Several challenges remain before clinical deployment The ensemble keeps six models in memory, which uses more storage than a single model. Model compression could reduce this. The system would also need to work with images from real dermoscopy devices, which may look different from the ISIC images used here. A tested interface would also be needed, so the dermatologist still makes the final decisions.

Table 9. Performance Comparison of the Proposed Framework with Previous Studies

Reference	Methodology	Accuracy
[12]	SeReNeXt-50 + EfficientNetB0 + EfficientNetB1 (MSM-CNN)	86.20%
[28]	ResNet50 + Xception + EfficientNetB0 + metadata	93.20%
[31]	CNN + SVM	92.00%
[35]	EfficientNetB3 + XGBoost (image + clinical data)	78.00%
[36]	ConvNeXtV2 + Separable Self-Attention	93.60%
[37]	Wavelet Decomposition + EfficientNet	92.20%
[45]	VGG16 (best of ML/DL comparison)	88.00%
Proposed	**Ensemble (EfficientNetB3 + Xception)**	**95.22%**

Table 9 compares the proposed ensemble framework with several previously published skin cancer classification methods. The proposed model achieved the highest classification accuracy of 95.22% among the selected methods included in this comparison. These results demonstrate the effectiveness of the proposed framework, which combines dataset-level fusion with ensemble learning. The fused dataset increases the diversity of the training data. The ensemble of Xception and EfficientNetB3 further improves the overall classification performance. As a result, the proposed framework provides reliable predictions for unseen skin lesion images. The performance of the system is better because it uses fusion and ensemble learning together. The fused datasets provide more diverse training samples. The ensemble then combines the predictions of EfficientNetB3 and Xception to improve the final classification results. These findings suggest that the proposed framework can serve as a reliable computer-aided tool to support skin cancer screening.

Table 8 already compares EfficientNetB3 alone, Xception alone, and the ensemble. All three were tested on the same test set. This shows the direct benefits of combining both models. Accuracy rose from 94.90 - 95.01% to 95.22%. A deeper study was not done here. It would remove SMOTE, Mixup, and dataset fusion one at a time. This is left for future work. A statistical test between the ensemble and single models was also not done. This test needs matched predictions on each test image. These were not saved separately during training. This is planned as a next step. Testing on another public dataset, such as HAM10000, is also planned. This is a two-class problem. Because of this, sensitivity and specificity already show how the model performs on each class separately.

Conclusion:

This study presents an ensemble transfer learning framework for binary skin cancer classification. The framework uses dermoscopic images from the ISIC 2016 and ISIC 2017 datasets. Two transfer learning models were used: EfficientNetB3 and Xception. Their predictions were combined using ensemble averaging. The proposed ensemble achieved a test accuracy of 95.22%. It outperformed both individual models. Xception alone scored 95.01%. EfficientNetB3 alone scored 94.90%. The ROC-AUC score of 0.9878 confirms that the ensemble gives a strong and reliable separation between benign and malignant cases. This study makes three primary contributions. First, two public datasets were combined to increase training diversity. Second, SMOTE and Mixup were used together to handle class imbalance and improve generalization. Third, EfficientNetB3 and Xception were combined through ensemble averaging, and this combination outperformed both models individually. Despite these encouraging results, several limitations should be acknowledged. The dataset remains moderate in size, even after combining ISIC 2016 and ISIC 2017. This is still small for deep learning models. The study only performs binary classification. It does not separate different types of malignant lesions, such as melanoma or basal cell carcinoma. The model was tested on dermoscopic images only. Its performance on clinical smartphone images has not yet been evaluated. No patient metadata was used, which could have helped improve accuracy further. The ensemble also relies on simple equal-weighted averaging. More advanced ensemble strategies were not explored. SMOTE creates new malignant samples by mixing existing ones. These new samples may not fully match the range of real malignant cases. This could make the model learn patterns that are smoother than real cases. Also, ISIC 2016 and ISIC 2017 were collected using different devices and settings. Combining them may cause a small shift between the two sources, even though both are dermoscopic datasets. Common steps like resizing and colour conversion were used to reduce this risk, but it was not measured directly. Future work can further improve this framework in several ways. The model can be extended to multi-class classification. This would allow it to identify the exact type of skin cancer, not just benign or malignant. The model can also be tested on clinical, non-dermoscopic images. This will help evaluate its performance in real-world mobile and telemedicine settings. Patient metadata, such as age and lesion location, may also be incorporated to improve accuracy. Larger and more diverse datasets could further improve reliability across different populations. More advanced ensemble strategies, such as weighted or learned ensembles, can also be explored. Explainable AI techniques, such as Grad-CAM, can also be incorporated to highlight the image regions that influenced the model's predictions. This can help dermatologists better understand and trust the system. Future work should also include validation with practicing dermatologists in real clinical settings. Next, we will test the trained models on a new dataset, like HAM10000. We will use McNemar's test to compare the ensemble model with the individual models. A deeper study that removes SMOTE, Mixup, and dataset fusion one at a time is also planned, using the same trained pipeline.

Acknowledgment:

The authors acknowledge the International Skin Imaging Collaboration (ISIC) for providing public access to the ISIC 2016 and ISIC 2017 datasets used in this study.

References:

[1] B. F. . ung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *CA Cancer J Clin*, vol. 71, no. 3, pp. 209–249, 2021, doi: 10.3322/caac.21660.

[2] S. J. . Fried L, Tan A, Bajaj S, Liebman TN, Polsky D, "Technological advances for the detection of melanoma: Advances in diagnostic techniques," *J Am Acad Dermatol*, vol. 83, no. 4, pp. 983–992, doi: 10.1016/j.jaad.2020.03.121.

[3] M. Janda *et al.*, "Early detection of melanoma: a consensus report from the Australian Skin and Skin Cancer Research Centre Melanoma Screening Summit," *Aust. N. Z. J. Public Health*, vol. 44, no. 2, pp. 111–115, Apr. 2020, doi: 10.1111/1753-6405.12972.

[4] H. S. . Goyal M, Knackstedt T, Yan S, "Artificial intelligence-based image classification methods for diagnosis of skin cancer: Challenges and opportunities," *Comput Biol Med*, doi: 10.1016/j.compbiomed.2020.104065.

[5] D. K. . Davis S, Piggott C, Lyon C, "Effectiveness of dermoscopy in skin cancer diagnosis," *Can Fam Physician*, vol. 66, no. 10, pp. 739–740, [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC7571636/

[6] K. H. . Sinz C, Tschandl P, Rosendahl C, Akay BN, Argenziano G, Blum A, Braun RP, Cabo H, Gourhant JY, Kreusch J, Lallas A, Lapins J, Marghoob AA, Menzies SW, Paoli J, Rabinovitz HS, Rinner C, Scope A, Soyer HP, Thomas L, Zalaudek I, "Accuracy of dermatoscopy for the diagnosis of nonpigmented cancers of the skin.," *J Am Acad Dermatol*, vol. 77, no. 6, pp. 1100–1109, doi: 10.1016/j.jaad.2017.07.022.

[7] J. M. Jung, J. Y. Cho, W. J. Lee, S. E. Chang, M. W. Lee, and C. H. Won, "Emerging Minimally Invasive Technologies for the Detection of Skin Cancer," *J. Pers. Med.*, vol. 11, no. 10, Oct. 2021, doi: 10.3390/JPM11100951.

[8] T. J. Brinker *et al.*, "A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task," *Eur. J. Cancer*, vol. 111, pp. 148–154, Apr. 2019, doi: 10.1016/j.ejca.2019.02.005.

[9] H. A. Haenssle *et al.*, "Man against Machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists," *Ann. Oncol.*, vol. 29, no. 8, pp. 1836–1842, Aug. 2018, doi: 10.1093/annonc/mdy166.

[10] M. A. Kassem, K. M. Hosny and M. M. Fouad ., "Skin Lesions Classification Into Eight Classes for ISIC 2019 Using Deep Convolutional Neural Network and Transfer Learning," *IEEE Access*, vol. 8, pp. 114822–114832, 2020, doi: 10.1109/ACCESS.2020.3003890.

[11] N. Tajbakhsh *et al.*, "Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?," *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1299–1312, May 2016, doi: 10.1109/TMI.2016.2535302.

[12] E. I. . Mahbod A, Schaefer G, Wang C, Dorffner G, Ecker R, "Transfer learning using a multi-scale and multi-network ensemble for skin lesion classification," *Comput Methods Programs Biomed*, 2020, doi: 10.1016/j.cmpb.2020.105475.

[13] A. F. 5 Munir Ahmad 1 , Majed Alfayad 2 , Shabib Aftab 1, 3 , Muhammad Adnan Khan 4, * and Bilal Shoaib 6 , Mohammad Sh. Daoud 7 and Nouh Sabri Elmitwally., "Data and Machine Learning Fusion Architecture for Cardiovascular Disease Prediction," *Mater. Contin.*, vol. 69, no. 2, doi: http://dx.doi.org/10.32604/cmc.2021.019013.

[14] Samia Akhtar,Shabib Aftab., "A Framework for Diabetic Retinopathy Detection using Transfer Learning and Data Fusion," *Res. gate*, vol. 16, no. 6, pp. 61–73, doi: 10.5815/ijitcs.2024.06.05.

[15] Samia Akhtar,Shabib Aftab., "Diabetic Retinopathy Severity Classification Using Data Fusion and Ensemble Transfer Learning," *J. Softw. Eng. Appl.*, vol. 18, no. 1, pp. 1–23, doi: 10.4236/jsea.2025.181001.

[16] Aftab, S., Abbas, S., Ghazal, T. M., Ahmad, M., Hamadi, H. A., Yeun, C. Y., & Khan, M. A ., "A Cloud-Based Software Defect Prediction System Using Data and Decision-Level Machine Learning Fusion," *Mathematics*, vol. 11, no. 3, p. 632, 2023, doi: https://doi.org/10.3390/math11030632.

[17] C. R. & H. K. Philipp Tschandl, "The HAM10000 dataset, a large collection of multi-

source dermatoscopic images of common pigmented skin lesions," *Sci. Data Vol.*, vol. 5, no. 180161, 2018, doi: https://doi.org/10.1038/sdata.2018.161.

[18] A. G. C. Pacheco *et al.*, "PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones," *Data Br.*, vol. 32, p. 106221, Oct. 2020, doi: 10.1016/J.DIB.2020.106221.

[19] David Gutman, Noel C. F. Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, Allan Halpern., "Skin Lesion Analysis toward Melanoma Detection: A Challenge at the International Symposium on Biomedical Imaging (ISBI) 2016, hosted by the International Skin Imaging Collaboration (ISIC)," *arXiv:1605.01397*, doi: https://doi.org/10.48550/arXiv.1605.01397.

[20] N. C. F. Codella *et al.*, "Skin Lesion Analysis Toward Melanoma Detection: A Challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), Hosted by the International Skin Imaging Collaboration (ISIC)," *Proc. - Int. Symp. Biomed. Imaging*, vol. 2018-April, pp. 168–172, Oct. 2017, doi: 10.1109/ISBI.2018.8363547.

[21] L. Yu, H. Chen, Q. Dou, J. Qin and P. -A. Heng., "Automated Melanoma Recognition in Dermoscopy Images via Very Deep Residual Networks," *IEEE Trans. Med. Imaging*, vol. 36, no. 4, pp. 994–1004, 2017, doi: 10.1109/TMI.2016.2642839.

[22] T. G. Dietterich, "Ensemble methods in machine learning," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 1857 LNCS, pp. 1–15, 2000, doi: 10.1007/3-540-45014-9_1/SAVE-RESEARCH.

[23] Q. V. Le Mingxing Tan, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," *arXiv:1905.11946*, 2019, doi: https://doi.org/10.48550/arXiv.1905.11946.

[24] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-January, pp. 1800–1807, Nov. 2017, doi: 10.1109/CVPR.2017.195.

[25] N. V. Chawla, K. W. Bowyer, "SMOTE: Synthetic Minority Over-sampling Technique," *arXiv:1106.1813*, 2011, [Online]. Available: https://arxiv.org/abs/1106.1813

[26] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, Dec. 2019, doi: 10.1186/S40537-019-0197-0/FIGURES/33.

[27] Halder, A., Dalal, A., Gharami, S. et al., "A fuzzy rank-based deep ensemble methodology for multi-class skin cancer classification," *Sci Rep*, vol. 15, no. 6268, 2025, doi: https://doi.org/10.1038/s41598-025-90423-3.

[28] N. L. B. & N. A. H. Mayar Ashraf Saeed, Yasmine M. Afify and Scientific Reports volume ., "Multimodal deep learning ensemble framework for skin cancer detection," *Sci Rep*, vol. 15, no. 45660, doi: https://doi.org/10.1038/s41598-025-30534-z.

[29] Dorathi Jayaseeli, J.D., Briskilal, J., Fancy, C. et al., "An intelligent framework for skin cancer detection and classification using fusion of Squeeze-Excitation-DenseNet with metaheuristic-driven ensemble deep learning models," *Sci. Rep.*, vol. 15, no. 7425, 2025, doi: https://doi.org/10.1038/s41598-025-92293-1.

[30] Priya Natha, Sivarama Prasad Tera, Ravikumar Chinthaginjala, Safia Obaidur Rab, C. Venkata Narasimhulu & Tae Hoon Kim., "Boosting skin cancer diagnosis accuracy with an ensemble approach," *Sci. Rep.*, doi: https://www.nature.com/articles/s41598-024-84864-5.

[31] A. Asaduzzaman, C. C. Thompson, F. N. Sibai, and M. J. Uddin, "Application of ensemble learning models in computer-aided diagnosis of skin diseases," *Neural Comput. Appl. 2025 3721*, vol. 37, no. 21, pp. 16735–16751, Jun. 2025, doi:

10.1007/S00521-025-11336-W.

[32] K. Matsunaga, A. Hamada, A. Minagawa, and H. Koga, "Image Classification of Melanoma, Nevus and Seborrheic Keratosis by Deep Neural Network Ensemble," Mar. 2017, Accessed: Aug. 15, 2026. [Online]. Available: https://arxiv.org/pdf/1703.03108

[33] Yang, G.; Luo, S.; Greer, P., "Boosting Skin Cancer Classification: A Multi-Scale Attention and Ensemble Approach with Vision Transformers," *Sensors*, vol. 25, no. 8, 2025, doi: https://doi.org/10.3390/s25082479.

[34] Vijaya, J., Roy, A., Gyanchandani, B. et al., "A novel particle swarm optimization-based ensemble efficientnet learning model for imbalanced image classification in medical diagnosis: case study of melanoma skin cancer prediction," *Netw Model Anal Heal. Inf. Bioinforma*, vol. 14, no. 119, 2025, doi: https://doi.org/10.1007/s13721-025-00619-w.

[35] Khan, Irfan Ullah, Aslam, Nida, Anwar, Talha, Aljameel, Sumayh S., Ullah, Mohib, Khan, Rafiullah, Rehman, Abdul, Akhtar, Nadeem., "Remote diagnosis and triaging model for skin cancer using EfficientNet and extreme gradient boosting," *Complexity*, p. 13, 2021, doi: https://doi.org/10.1155/2021/5591614.

[36] Burhanettin Ozdemir & Ishak Pacal ., "A robust deep learning framework for multiclass skin cancer classification," *Sci. Reports Vol.*, vol. 15, 2025, doi: https://doi.org/10.1038/s41598-025-89230-7.

[37] A. I. Aboulmira A, Hrimech H, Lachgar M, Hanine M, Garcia CO, Mezquita GM, "Hybrid model with wavelet decomposition and EfficientNet for accurate skin cancer classification," *J Cancer*, vol. 16, no. 2, pp. 506–520, 2016, doi: 10.7150/jca.101574.

[38] Krohling, A. G. C. Pacheco and R. A., "An attention-based mechanism to combine images and metadata in deep learning models applied to skin cancer classification," *IEEE J. Biomed. Heal. Informatics,* vol. 25, no. 9, pp. 3554–3563, 2021, doi: 10.1109/JBHI.2021.3062002.

[39] C. J. and J. Y. (2022) A. study on skin tumor classification based on dense convolutional networks with fused metadata. Yin W, Huang J, "A study on skin tumor classification based on dense convolutional networks with fused metadata," *Front. Oncol.*, 2022, doi: https://doi.org/10.3389/fonc.2022.989894.

[40] D. L. Ha, S. Jeong, D. Woo, J. Cho, and W. J. Lee, "Integrating Anatomical Site Information Into Vision Transformer Models for Skin Cancer Classification," *Ann. Dermatol.*, vol. 37, no. 5, pp. 314–316, Oct. 2025, doi: 10.5021/AD.24.181.

[41] Leandro M. de Lima, Renato A. Krohling., "Exploring Advances in Transformers and CNN for Skin Lesion Diagnosis on Small Datasets," *IEEE Access*, 2022, doi: https://doi.org/10.1007/978-3-031-21689-3_21.

[42] M. L. Battle, A. Atapour-Abarghouei and A. S. McGough., ""Siamese Neural Networks for Skin Cancer Classification and New Class Detection using Clinical and Dermoscopic Image Datasets," *EEE Int. Conf. Big Data*, pp. 4346–4355, 2022, doi: 10.1109/BigData55660.2022.10020820.

[43] M. Azeem, K. Kiani, T. Mansouri, and N. Topping, "SkinLesNet: Classification of Skin Lesions and Detection of Melanoma Cancer Using a Novel Multi-Layer Deep Convolutional Neural Network," *Cancers 2024, Vol. 16, Page 108*, vol. 16, no. 1, p. 108, Dec. 2023, doi: 10.3390/CANCERS16010108.

[44] Bechelli S, Delhommelle J., "Machine learning and deep learning algorithms for skin cancer classification from dermoscopic images," *Bioengineering*, vol. 9, no. 3, p. 97, 2022, doi: 10.3390/bioengineering9030097.

[45] G. M. S. Himel, M. M. Islam, K. A. Al-Aff, S. I. Karim, and M. K. U. Sikder, "Skin Cancer Segmentation and Classification Using Vision Transformer for Automatic

Analysis in Dermatoscopy-Based Noninvasive Digital System," *Int. J. Biomed. Imaging*, vol. 2024, 2024, doi: 10.1155/2024/3022192.

[46] H. T. Halawani, E. M. Senan, Y. Asiri, I. Abunadi, A. M. Mashraqi, and E. A. Alshari, "Enhanced early skin cancer detection through fusion of vision transformer and CNN features using hybrid attention of EViT-Dens169," *Sci. Reports 2025 151*, vol. 15, no. 1, pp. 34776-, Oct. 2025, doi: 10.1038/s41598-025-18570-1.

[47] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2017-October, pp. 618–626, Dec. 2017, doi: 10.1109/ICCV.2017.74.

[48] A. M. Fayyaz, S. J. Abdulkadir, N. Talpur, S. M. Al-Selwi, S. U. Hassan, and E. H. Sumiea, "Grad-CAM (Gradient-weighted Class Activation Mapping): A systematic literature review," *Comput. Biol. Med.*, vol. 198, p. 111200, Nov. 2025, doi: 10.1016/J.COMPBIOMED.2025.111200.

[49] S. Yan *et al.*, "A multimodal vision foundation model for clinical dermatology," *Nat. Med. 2025 318*, vol. 31, no. 8, pp. 2691–2702, Jun. 2025, doi: 10.1038/s41591-025-03747-y.

[50] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond Empirical Risk Minimization," *6th Int. Conf. Learn. Represent. ICLR 2018 - Conf. Track Proc.*, Oct. 2017, Accessed: Aug. 15, 2026. [Online]. Available: https://arxiv.org/pdf/1710.09412