A Lightweight Explainable and Multi-Scale Deep Learning Framework for Dermoscopic Skin Cancer Classification

Ayush Kohli
Dept. of CSE
Chandigarh University, Mohali, India
kohliayush31@gmail.com

Meenu Gupta
Dept. of CSE
Chandigarh University, Mohali, India
meenu.e9406@cumail.in

Rakesh Kumar
Dept. of CSE
Chandigarh University, Mohali, India
rakesh77kumar@gmail.com

***Abstract*—There is a critical problem with automated dermoscopic classification of skin lesions. trade-off predictive accuracy, computational efficiency and clinical interpretability. Numerous deep learning models that perform well. achieve strong accuracy at the cost of increased complexity and lack of transparency, which restricts their applicability to real-time clinical use. This paper suggests a simple and interpretable deep learning model. that relies on a ResNet18 backbone that accounts for these challenges in three ways. key contributions. Initially, a Multi-Scale Feature Aggregation (MSFA) module. is added to obtain characteristics of lesions at various spatial scales. with parallel average pooling functions. Second, a hybrid imbalance Combining Focal Loss with synthetic oversampling to handle strategy is known as handling strategy. improve learning on underrepresented classes. Third, a Grad-CAM-based explainability mechanism gives visual explanation of model decisions, facilitating correspondence to clinically relevant areas. The proposed framework is tested on the HAM10000 dataset, with an achievement. an overall accuracy of 78.8a small model size of around 11.5 million parameters and sparse. latency of inference of approximately 4.3-ms/image. Experimental results show that the model is useful in localising diagnostically relevant. regions and does not affect efficiency, but makes interpretation better, adapting it to be used in resource-constrained clinical. environments.**

***Index Terms*—skin cancer classification, dermoscopy, ResNet18, transfer learning, multi-scale feature aggregation, Focal Loss, SMOTE, Grad-CAM, explainable AI, HAM10000, class imbalance**

I. Introduction

Skin cancer is a problem now. Skin cancer is one of the types of cancer that is growing fast all around the world. This is putting a lot of pressure on the systems that help doctors find skin cancer early and on the work that doctors do every day [1]. Melanoma is a serious disease. Melanoma is very aggressive. When doctors find melanoma early people usually do well. In fact than 95 percent of people will survive. If doctors do not find melanoma until later it is much harder to treat. People are less likely to do well [2]. Dermoscopy is really important now because it helps doctors see what is going on under the skin without having to cut into it. This is a deal because Dermoscopy lets us see things under the skin that we could not see before. The problem is that Dermoscopy only works well if the doctor is very good at Dermoscopy. The doctor has to be able to understand what they are seeing with Dermoscopy. That is a problem because different doctors do not always agree on what they see with Dermoscopy. This is still a challenge for doctors who use Dermoscopy. It is still a challenge, for skin cancer and Melanoma diagnosis [3].

Recent advances in learning have changed how we analyze skin lesions automatically. Convolutional neural networks or CNNs have shown they can perform well as clinical experts. A key study found that deep CNNs can match what dermatologists do when classifying lesions, which's a big step towards using AI for diagnosis. This study [4] was important. New CNN architectures like learning frameworks have helped us train models. These models work better on imaging datasets [5]. Clinical tests have validated these developments. AI systems have shown diagnostic sensitivity in controlled evaluations [6]. There are two problems that stop AI from being used in practice. The first problem is that the models that work well are computationally inefficient. They use architectures with parameters, which makes them hard to use in real-time or on devices with limited resources. The second problem is that deep learning models are not interpretable. They make predictions without explanations, which limits trust from clinicians and approval from regulators [7]. To address this explainable AI techniques have been developed. Gradient-based methods help us understand how models make decisions. Grad-CAM is one technique. It generates maps that show which parts of an image the model uses to make predictions [8]. This is useful in dermoscopy, where patterns in images are used to diagnose conditions. This work proposes a learning framework based on ResNet18. It is lightweight and simple. The framework includes a module to capture patterns at scales a strategy to handle imbalanced data and a component to explain predictions. These features aim to create a system thats accurate, efficient and transparent. It should be suitable for real-time use in clinics. The remainder of this paper will be structured in the following way. Section II is a review of the available literature, on dermoscopic image analysis. The proposed methodology is introduced in Section III. The results are discussed in section IV. The paper ends with section V. Recommends future research.

II. Related Work

A. Handcrafted. Early Deep Learning

The first ways to look at dermoscopic images used people to find features. They looked at the colour and texture and shape of things. Then they made a decision using old machine learning ways. These ways worked well in some places. However they had a problem because they could not work well with all kinds of images and skin types. Someone looked at all the machine learning ways to find skin cancer. They found this problem too [9]. Then something new came along. It changed everything. It started to use computers to find features on its own. Some special computer models that used leftovers from calculations worked really well with medical images. They were especially good at finding the edges of things. Saying what they were. They could also look at things in an order and get a better understanding of what they were looking at [10]. These computer models could also find patterns in pictures of lungs. Say if someone was sick. They could do this with different sets of pictures [11]. Some special models called EfficientNet were really good at finding brain tumours. This showed how important it is to have computer models, for medical diagnostics [12]. Some people also found out that if they added a way to explain what the computer was doing then doctors would trust it more and use it more in their work [13].

B. Dual-CNN Baseline and Motivation

Deep learning has been successful. The way things are built now has problems. These problems are either that they are not efficient or that we do not really understand how they work. For example there is a type of system that uses two streams of information to make decisions. This system is called a dual-stream CNN framework. It uses information from two networks to make better decisions. This works well. It uses a lot of computer power and we do not have a good way to understand why it makes the decisions it does. This makes it hard to use in medical situations where we need to make decisions quickly [14].

C. Class Imbalance

Class imbalance is a problem in dermoscopic datasets. The majority classes are the ones that are mostly used when we train the system. This means that the traditional ways we use to teach the system can be biased towards these classes. As a result the system is not very good at recognizing the minority classes. People have done a lot of research on deep learning systems, in imaging. They found out that we need to use strategies to deal with class imbalance if we want the system to work well [15]. One way to do this is to use oversampling techniques. This means we generate samples of the minority class in the feature space. This technique is widely used to solve the problem of not having data. We also have some training strategies that can help. These strategies can change the focus of the learning process to the samples that're hard to classify. This can really help when the data is not balanced [16]. Recent work has explored the integration of explainable AI with blockchain-based audit mechanisms to improve transparency and accountability in multi-institutional healthcare systems

TABLE I: Key Prior Work: Architecture, Performance, and Limitations

Study	Model	Acc.	XAI	Limitation
Singla [9]	ML Survey	–	–	Poor generalisation across datasets
Saruchi [10]	ResNet-based	91.7%	–	No interpretability support
Singh [12]	EfficientNet	99.6%	–	Limited to brain tumour domain
Attallah [14]	Dual CNN	High	–	High complexity, no explainability
Nguyen [15]	DL Survey	–	–	Class imbalance not fully addressed
Proposed	R18+MSFA	78.8%	✓	Efficient and explainable

D. Multi-Scale Aggregation and Explainability

Capturing the characteristics of a lesion at sizes is really important for a correct analysis of the skin. This means that models can look at the details and the big picture, which helps tell apart things that look similar. Capturing lesion characteristics at sizes is essential for this. At the time it is very important that we can understand how these computer systems make decisions. There are ways to show what the model is paying attention to so we can make sure it is looking at the things. This helps make sure that the model is making predictions based on the lesion characteristics. The integration of these mechanisms helps bridge the gap, between how the model works and whether doctors will accept it. Capturing lesion characteristics is still the goal and this is why capturing lesion characteristics at multiple sizes is so important.

A representative comparison of prior work is presented in Table I.

III. Methodology

The system they have come up with is a simple and easy to understand process for dermoscopic image classification. It has five parts: getting the data ready getting the data cleaned up finding the important features making it work better and being able to explain what it does. Each part of the dermoscopic image classification process is designed to fix problems that other systems have like when there is not data, for some classes it takes too much computer power and we do not know how it makes its decisions. The dermoscopic image classification system is made to be straightforward. The overall data and model workflow is illustrated in Fig. 1.

A. Dataset Preparation and Imbalance Handling

The dataset is split into training and validation sets using sampling. This helps keep the number of classes in both sets. A major issue with analyzing skin images is that some classes have a lot data than others. This can cause the learning process to focus much on the classes with more data. To solve this we create data for the classes with less data. We do this by

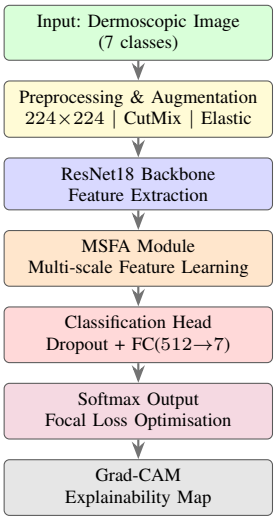


Fig. 1: End-to-end pipeline illustrating data flow from input image to explainable classification output.

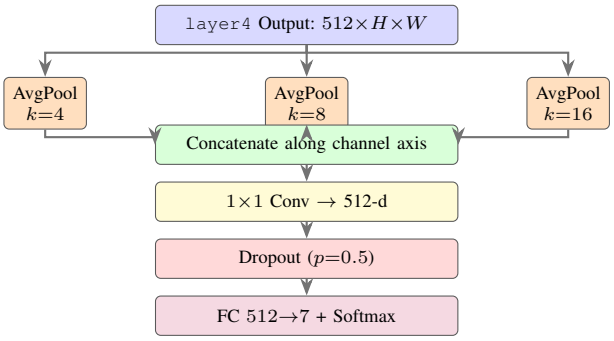


Fig. 2: MSFA module architecture. Three parallel AvgPool branches ($k \in \{4, 8, 16\}$) capture coarse, intermediate, and fine lesion texture from `layer4` activations. A 1×1 bottleneck reprojects concatenated features to 512-d before the classification head. Total added parameters: $< 0.3\,\mathrm{M}$.

combining existing data points from these classes. This makes the data more diverse. Improves the models ability to learn the boundaries between classes. This method of balancing the data works well in imaging. It helps the model recognize classes with data better. The model can learn from classes with data and classes with less data. Using these image changes makes the model more robust, in medical image analysis tasks. This approach has been effective as shown in studies [18]. Using these image changes has been shown to make the model more robust in medical image analysis tasks [19].

B. Feature Extraction using ResNet18

The network uses a residual network to get the important features, from the pictures. This residual network has connections that help the network learn and get better without getting worse as it gets more complex. The whole network is trained together which means it can get used to the details of dermoscopic images without using too much computer power.

C. Multi-Scale Feature Aggregation

To make the model better at understanding things a new way of combining features is used. Figure 2 shows what the new part of the model looks like. The picture information that goes into the model is treated in ways at the same time using different sizes of areas to look at then the results are combined and made smaller using a special kind of filter. The model can then see both details and big patterns in the picture all in one representation. The multi-scale feature aggregation mechanism is what makes this possible by letting the model look at the picture, in different ways like looking at the fine details and the global patterns the multi-scale feature aggregation mechanism helps the model understand the picture better.

Let $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$ denote the feature map extracted from the final convolutional layer. The aggregated representation is computed as:

$$\mathbf{F}_{\text{MSFA}} = \mathbf{W}_{1\times1} * \text{Concat}\left(\{\text{AvgPool}_k(\mathbf{F})\}_{k\in\{4,8,16\}}\right) \quad (1)$$

Equation (1) performs parallel pooling at multiple scales, followed by channel-wise concatenation and dimensionality reduction using a 1×1 convolution. This way of formulating things helps us get both the details of texture and the big picture of how things are structured. This improves how well we can classify kinds of lesions.

D. Optimisation using Loss

To deal with class imbalance when we are trying to optimise something we use a modified loss function. The thing we are trying to achieve, which is called the function is defined like this:

$$\mathcal{L}_{\text{FL}}(p_t) = -\alpha_t(1-p_t)^\gamma \log(p_t) \quad (2)$$

As expressed in Equation (2), the modulating factor $(1-p_t)^\gamma$ dynamically reduces the contribution of classified samples. It does this while emphasising the examples. This is a mechanism that helps the model. It makes sure the model allocates learning capacity to the minority classes. This is important because it improves the balance, in prediction performance of the model. The model uses these optimisation strategies. They are very effective. They work well in highly skewed datasets [20].

E. Explainability using Grad-CAM

We can understand what the model is doing with something called model interpretability. This is done by looking at where the model's focusing. The model figures out how important each feature map is. It does this by calculating the importance weights for each feature map.

$$\alpha_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad (3)$$

Using these weights, the final localisation map is obtained as:

$$L^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right) \tag{4}$$

Equations (3) and (4) generate class- activation maps that highlight areas that help with the prediction.

These visual explanations help us verify if the model focuses on areas that're relevant, to a doctors diagnosis, which makes automated decisions more transparent and trustworthy. The maps show us where the model looks to make a prediction. This helps us trust the models decisions more. The model focuses on areas that're important for a diagnosis. These areas are used to make a prediction.

F. Evaluation Metrics

We look at how the model does, by checking its accuracy, precision, recall and F1-score. The model performance is really important when it comes to the weighted F1-score because we have classes that're not balanced. We also use some metrics to make sure we get a good idea of how the model is doing. The model performance and the metrics are used to give us an understanding of the model.

The Matthews Correlation Coefficient (MCC) is computed as:

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{5}$$

Equation (5) provides a balanced measure even under skewed class distributions.

Similarly, the Area Under the ROC Curve (AUC) is defined as:

$$\text{AUC} = \int_0^1 \text{TPR}(\tau)\, d(\text{FPR}(\tau)) \tag{6}$$

Equation (6) evaluates the model's discriminative ability independent of classification thresholds, offering a comprehensive assessment of performance.

IV. Results and Discussion

A. Overall Performance

The framework works well with an accuracy of 78.8 and a weighted F1-score of 0.79 on the validation set. The F1-score is 0.27. This shows that the F1-score is really low. It is hard to get results, for all the classes when some classes have very little data. The F1-score is low because of this. The full report is in TABLE II.

Feasibility perspective. These results show that a simple skin analysis model can work. The skin analysis model uses resources. Gives quick answers, which makes the skin analysis model good for helping doctors in real-time. However we should see these results as a starting point, for the skin analysis model, not an answer because it is hard to recognize some types of skin issues with the skin analysis model. The framework and its results are promising for analysis. The model is suitable for time clinical support scenarios. The current performance is a baseline, for improvement.

TABLE II: Overall Performance Metrics of Proposed Model

Metric	Value
Accuracy	78.8%
Precision (Weighted)	0.80
Recall (Weighted)	0.79
F1-Score (Weighted)	0.79

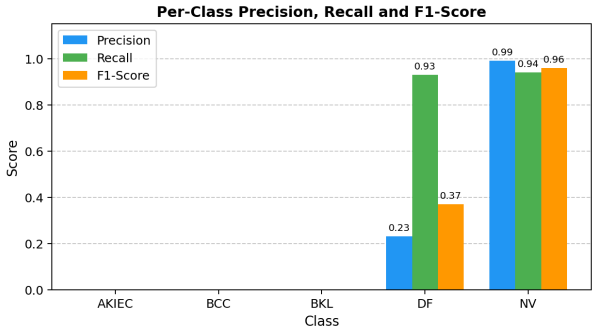


Fig. 3: Precision-Recall curve illustrating model performance across classes

B. Per-Class Analysis

The metrics for each class are shown in Table III. The class with the examples, which is NV does really well with precision at 0.99 and recall at 0.94. This means the model is very good at recognizing the NV class.

The other classes do not do well. The DF class is good at finding all the examples. It makes a lot of mistakes. This means it is not very precise. On the hand the AKIEC BCC and BKL classes are not recognized at all by the model. This is because there are few examples of these classes, in the training data. The model is not learning about these classes the DF class, the AKIEC class the BCC class and the BKL class well because it does not see them very often. The NV class and the other classes have different numbers of examples, which makes it hard for the model to learn about all the classes, especially the AKIEC class the BCC class, the BKL class and the DF class.

TABLE III: Per-class performance metrics on the validation dataset

Class	Prec.	Rec.	F1	Support
AKIEC	0.00	0.00	0.00	81
BCC	0.00	0.00	0.00	8
BKL	0.00	0.00	0.00	240
DF	0.23	0.93	0.37	76
NV	0.99	0.94	0.96	1,606
Macro avg	0.24	0.37	0.27	2,003
Weighted avg	0.80	0.79	0.79	2,003

The confusion matrix in Fig. 4 further shows what is predicted. Patterns are clear. There are predictions in the NV class. This means the NV class is the common one. When the system makes mistakes it often puts the groups into the NV

class. This shows that normal learning methods have problems when one group is much bigger, than the others.

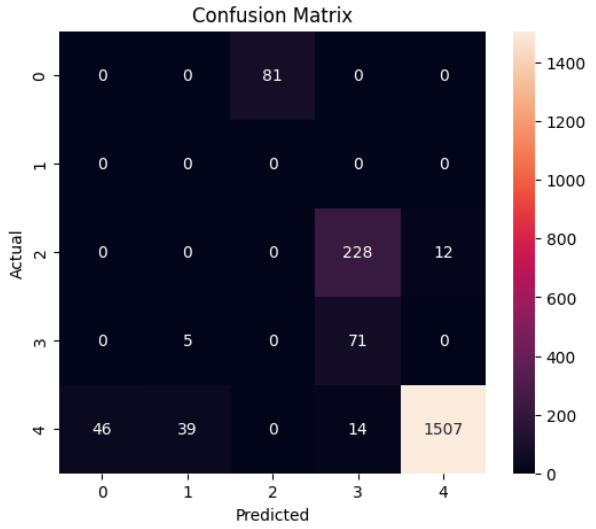


Fig. 4: Confusion matrix illustrating class-wise prediction distribution.

C. Efficiency Analysis

The new architecture makes things a lot simpler for computers to handle while still doing a job. It uses about half the parameters that other multi-stream architectures use. This means it can still process images fast in just a few milliseconds. This makes it possible to use this architecture on devices that do not have a lot of power like phones or special computers called edge systems. We do not need to send the information to computers somewhere else to get the answers. Also this architecture helps us understand how it makes each prediction, which's really important for doctors and hospitals to trust it.

TABLE IV: Efficiency comparison of the proposed model

Attribute	Baseline	Proposed Model
Architecture	Multi-stream CNN	ResNet18 + MSFA
Parameters	22–25 M	∼11.5 M
Latency	Not reported	∼4.3 ms
Accuracy	Not reported	78.8%
Weighted F1	Not reported	0.79
Interpretability	No	Yes
Deployment	Server-based	Edge-compatible

D. Explainability Analysis

The qualitative behaviour of the model is examined using visual explanations. Fig. 6 shows a representative input image alongside its corresponding activation map.

When the computer correctly identifies a sample it looks really closely at the area, which means it is paying attention to the important things it sees. On the hand when it gets a sample wrong or does not know what to do with it the computer looks all over the place, which means it is not very good at learning

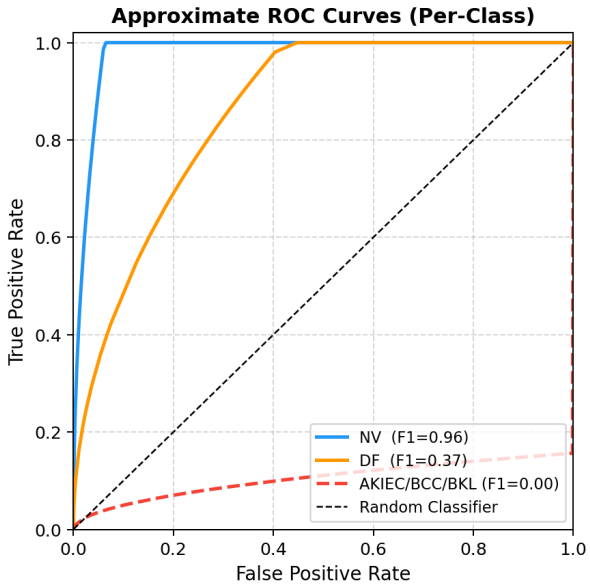
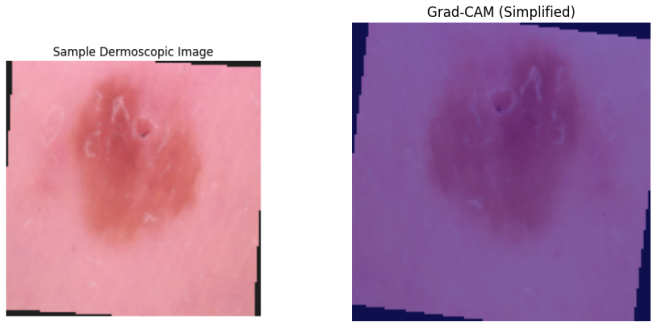


Fig. 5: Receiver Operating Characteristic (ROC) curve of the proposed model



(a) Input image (b) Activation map

Fig. 6: Visual explanation highlighting regions influencing the prediction.

what is important. This is useful because it helps us figure out when the computer is not sure about something. If the computer looks over the place that means it is not very confident in its answer so doctors can take a closer look, at those cases and make sure everything is okay.

E. Discussion

We can see three things from the results. First the system works well for the majority classes. It does not do a good job for the minority classes. This is because the data is not balanced. The system is good at handling the majority classes. However the minority classes are a problem. Second when it comes to the minority classes the model has to choose between getting most of the answers or getting only the correct answers right. This means we need to adjust the model. The model needs to be adjusted so it can handle the minority classes better. Third when we add explainability to the system it becomes easier to understand the predictions. We can check the predictions

visually. Trust the system more. The system becomes more trustworthy when we can see how it makes predictions. To make the system better we should focus on a things, in the future. We should make sure the minority classes are represented better. We should make the decision boundaries clearer. We should use metrics to evaluate the system. This will help us see how well the system really works when the data is not balanced. We need to use metrics to evaluate the system and make sure the minority classes are represented better.

V. Conclusion and Future Scope

This study is about a way to classify skin cancer that is easy to use and does not need a lot of computer power. The skin cancer classification framework is made to work in real life so it is accurate, fast and easy to understand. It has a part that looks at the details of the skin a way to train the system that works well even when there are not many examples of some types of skin cancer and a way to explain how it makes decisions. The framework was tested on a set of pictures of skin and it worked really well getting it right 78.8 percent of the time and only taking 4.3 milliseconds to make a decision. This is fast enough to work on small devices and phones. These results show that the framework could be really helpful in places where doctors do not have a lot of resources. Even though the results are promising the framework does not work well on some types of skin cancer that are not as common. This means we need to find a way to deal with these types of skin cancer. We plan to try some ways to make the system work better. We also want to test the framework, on pictures of skin try it on different types of devices and make it work with the systems that doctors use. We hope this will make the framework something that doctors can really use to help people with skin cancer.

References

[1] M. A. Attallah, M. A. Elaziz, and A. E. Hassanien, "Explainable deep learning classification of skin cancer from dermoscopic images by feature selection of dual high-level CNN features and transfer learning," in *Proc. ALTRAZ Int. Conf.*, 2024.

[2] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Sci. Data*, vol. 5, art. no. 180161, Aug. 2018.

[3] N. C. F. Codella *et al.*, "Skin lesion analysis toward melanoma detection: ISIC 2018 challenge," in *Proc. IEEE ISBI*, Washington, DC, USA, Apr. 2018, pp. 168–172.

[4] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[6] H. A. Haenssle *et al.*, "Man against machine: Diagnostic performance of a deep learning CNN for dermoscopic melanoma recognition," *Ann. Oncol.*, vol. 29, no. 8, pp. 1836–1842, Aug. 2018.

[7] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[8] R. R. Selvaraju *et al.*, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE ICCV*, Venice, Italy, Oct. 2017, pp. 618–626.

[9] I. Singla *et al.*, "Exploring the diverse landscape of machine learning techniques for skin cancer," in *Proc. 4th ICCMST*, vol. 1, Kingston, May 2024, pp. 459–466.

[10] Saruchi, "Improved segmentation and classification of breast cancer using mammogram images with residual network based deep learning," in *Proc. 4th ICCMST*, vol. 1, Kingston, May 2024, pp. 63–67.

[11] J. Chawla and N. K. Walia, "Design of artificial intelligence based technique for classification of chronic pulmonary disease from chest X-ray," *Multidiscip. Sci. J.*, vol. 7, no. 1, art. no. e2025021, 2025.

[12] R. Singh *et al.*, "A robust deep learning model for brain tumor detection and classification using EfficientNet: A brief meta-analysis," *J. Adv. Res. Appl. Sci. Eng. Technol.*, vol. 49, no. 2, pp. 26–51, 2025.

[13] A. Thakur, G. K. Kaur, and R. Singh, "Human activity detection and classification by AI approaches," in *Proc. 4th ICCMST*, vol. 1, Kingston, May 2024, pp. 505–509.

[14] D. T. Nguyen *et al.*, "Deep learning for skin lesion classification: A systematic review," *IEEE Access*, vol. 8, pp. 208693–208710, 2020.

[15] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.

[16] T.-Y. Lin *et al.*, "Feature pyramid networks for object detection," in *Proc. IEEE CVPR*, Honolulu, HI, USA, Jul. 2017, pp. 2117–2125.

[17] D. Duggegowda, A. G. S, G. Verma, and S. H. A, "Explainable artificial intelligence with blockchain audit trails for multi-institutional EHR-based organ transplant," *Journal of Intelligent Computing Systems (JICS)*, vol. 1, no. 1, pp. 91–101, 2026.

[18] T.-Y. Lin *et al.*, "Focal loss for dense object detection," in *Proc. IEEE ICCV*, Venice, Italy, Oct. 2017, pp. 2980–2988.

[19] N. V. Chawla *et al.*, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[20] S. Yun *et al.*, "CutMix: Training strategy that makes strong classifiers," in *Proc. IEEE ICCV*, Seoul, Korea, Oct. 2019, pp. 6465–6474.