AMERICAN INTERNATIONAL UNIVERSITY-BANGLADESH

Faculty of Engineering

Final Report Cover Page

Assignment Title:	Reducing Silent Failures in Medical Image Classification: An Exploratory Investigation		
Assignment No:	3	Date of Submission:	27 August 2023
Course Title:	COMPUTER VISION AND PATTERN RECOGNITION		
Course Code:	COE4234	Section:	C
Semester:	Summer 2022-23	Course Teacher:	DR. DEBAJYOTI KARMAKER

Declaration and Statement of Authorship:

1. I/we hold a copy of this Assignment/Case-Study, which can be produced if the original is lost/damaged.
2. This Assignment/Case-Study is my/our original work and no part of it has been copied from any other student's work or from any other source except where due acknowledgement is made.
3. No part of this Assignment/Case-Study has been written for me/us by any other person except where such collaborationhas been authorized by the concerned teacher and is clearly acknowledged in the assignment.
4. I/we have not previously submitted or currently submitting this work for any other course/unit.
5. This work may be reproduced, communicated, compared and archived for the purpose of detecting plagiarism.
6. I/we give permission for a copy of my/our marked work to be retained by the Faculty for review and comparison, including review by external examiners.
7. I/we understand thatPlagiarism is the presentation of the work, idea or creation of another person as though it is your own. It is a formofcheatingandisaveryseriousacademicoffencethatmayleadtoexpulsionfromtheUniversity. Plagiarized material can be drawn from, and presented in, written, graphic and visual form, including electronic data, and oral presentations. Plagiarism occurs when the origin of them arterial used is not appropriately cited.
8. I/we also understand that enabling plagiarism is the act of assisting or allowing another person to plagiarize or to copy my/our work.

* *Student(s) must complete all details except the faculty use part.*

** Please submit all assignments to your course teacher or the office of the concerned teacher.

Group Name/No.: 6

No	Name	ID	Program	Signature
1	MD. Sazib Ahmed	20-42076-1	BSc [CSE]	
2	Badhan Akter	20-42225-1	BSc [CSE]	
3	Masud Pervez	20-42656-1	BSc [CSE]	
4			Choose an item.	
5			Choose an item.	
6			Choose an item.	
7			Choose an item.	
8			Choose an item.	
9			Choose an item.	
10			Choose an item.	

Faculty use only		
FACULTYCOMMENTS	Marks Obtained	
	Total Marks	

Reducing Silent Failures in Medical Image Classification: An Exploratory Investigation

MD. Sazib Ahmed (20-42076-1)[a],
Masud Pervez (20-42656-1)[a],
Badhan Akter (20-42225-1)[a]

[a] *Department of Computer Sciences, American International University-Bangladesh*

Abstract

For secure clinical deployment, Silent failure identification and mitigation in image classification are of utmost importance. Finding hidden failures is essential for triggering medical examination and maintaining patient safety. Despite the importance, there isn't enough concrete data to support the use of modern confidence score systems, particularly in medical imaging, to identify classification errors. This paper tries to fill this gap by suggesting an improved approach for identifying classification of images errors. The suggested method makes use of multiclass classification framework and confident scores on skin cancer images from the ISIC medical imaging dataset. The findings show, however, that it is still difficult to identify failures with accuracy. The study's findings highlight the challenges involved in accurate failure detection by utilizing a multiclass classification model. In the course of these investigations, it is noteworthy that coordinated efforts to improve a weighted total combining numerous confidence scores did not improve any accuracy. These results make it clear that a revised framework with better confidence scoring functions is necessary to increase the ability to detect failures in the classification of medical images. Code available at: https://github.com/Sazib-Ahmed/CVPR_Paper.git

Keywords: ***Silent Failure, Confident Scoring Function, Deep learning, Computer Vision, Medical Image Classification, ResNet50, CNN, ISIC***

1. Introduction

Numerous research (Calisto et al., 2021; Calisto et al., 2022) have shown the emergence of AI systems as possible enhancers of medical procedures in recent years. These initiatives aimed to lessen the workload for medical professionals and improve patient outcomes. However, due to their inherent fallibility, AI systems must be acknowledged to make mistakes (Shamshirband et al., 2021). Therefore, it has become essential to build strong countermeasures to avoid any potential negative effects resulting from these failures.

Numerous studies have concentrated on enhancing the relationship between AI and human components to improve patient outcomes as a result of the realisation of the necessity for such preventative measures. This method goes beyond the use of AI as a standalone computational entity and includes the integration of interpretability tools or the use of AI as an additional reviewer (Calisto et al., 2022, 2021; Shamshirband et al., 2021). However, it is becoming clearer that even "human safeguards" are not entirely resistant to failure, since inaccurate AI forecasts have the potential to fool even knowledgeable human evaluators (Tschandl et al., 2020).

Because of this, the detection and mitigation of silent failures in the classification of images have become crucial issues in the context of secure clinical deployment. Finding hidden failures is essential for triggering the right medical evaluations and guaranteeing patient safety. Nevertheless, despite the significance of this task, there is a clear a lack of solid real-world proof demonstrating the effectiveness of modern confidence scoring methods after deployment, particularly in the complex field of medical imaging, in identifying classification errors. This study attempts to close this important gap by recommending a better strategy for identifying silent failure in medical image classification.

The proposed method, which applies a multiclass classification architecture and confident scoring function to a set of skin cancer images from the ISIC medical imaging dataset, presents a fresh viewpoint. The investigation's conclusions, however, highlight the ongoing difficulty in identifying failures accurately. It's interesting to note that even advanced methods widely recommended in literature on computer vision and machine learning constantly fall short of outperforming a SoftMax average. (Bungert et al., 2023)

These provocative findings make it clear that a redesigned framework with improved confidence scoring functions is required to improve the ability to accurately identify failures in the classification of medical images. This study addresses this urgent requirement, advancing secure healthcare practices and furthering the conversation in the fields of deep learning, computer vision, and medical image classification.

This introduction lays the groundwork for the paper's next chapters, which will explore the goals, methods, results, and implications of the research. The complexity of the suggested approach will be thoroughly examined in the pages that follow, and then the research results and implications for the field of medical imaging will be carefully examined.

2. LITERATURE REVIEW

2.1. IMAGE CLASSIFICATION

Within the discipline of computer vision, which is concerned with giving computers the ability to comprehend and interpret visual data, image classification is a significant assignment. Images are automatically categorized into predetermined classes or categories as part of image classification. In order to identify the most appropriate label or class to which an image belongs in this task, machine learning algorithms examine the visual features and patterns contained in the image.

To teach a computer to recognize and distinguish between numerous objects, situations, or concepts inside images is the aim of image classification. Numerous practical uses exist for this ability, including identifying objects in images, spotting irregularities in medical images, recognizing faces, differentiating between different vehicle types, and more.

Convolutional neural networks (CNNs) and other deep learning models have revolutionized image classification by allowing computers to learn complex patterns and representations directly from raw pixel data. These models may automatically learn significant features that distinguish one class from another since they are trained on enormous datasets of labelled images. Image classification serves as a fundamental building element for many computer vision applications because, once trained, these models can correctly classify new, unknown images into the correct groups.

2.2. MEDICAL IMAGE CLASSIFICATION

Image classification in the context of medical imaging refers to the automated classification of medical images according to their visual characteristics. Convolutional neural networks (CNNs), which were trained on significant datasets of labelled medical images, are one of the more sophisticated machine learning approaches used in this procedure. The machine learning models gain the ability to recognise specific patterns and minor traits that discriminate between various medical illnesses, tissues, or anomalies in the images through this training.

Medical image classification has broad and significant effects. It is crucial in supporting healthcare professionals by providing automated assistance in classifying various medical scenarios. For example, it helps radiologists and physicians diagnose conditions including tumours, bone fractures, lung problems, and heart abnormalities with more accuracy and efficiency. Additionally, the technology is excellent at spotting irregularities that would escape manual inspection, aiding in an early and accurate diagnosis.

By quickly detecting serious conditions that demand immediate attention, medical image classification also facilitates the prioritisation of patients. It provides information on the seriousness and location of conditions, guiding medical interventions and operations. The ability to track a patient's development over time is helpful in determining the effectiveness of the treatment and making informed improvements. Furthermore, this technology can be used in medical research to automate the study of large datasets and uncover novel findings.

However, given the comprehensive nature of medical images, it is crucial to recognise the challenges posed by medical image categorization. Maintaining high standards for patient care requires that classification algorithms be accurate and reliable. As the area develops, efforts are focused on the creation of robust and understandable models that benefit healthcare professionals and ultimately result in improved patient outcomes and expanded medical research capabilities.

2.3. CONFIDENCE SCORE FUNCTION

The certainty or confidence a model has in its predictions can be measured using a confidence score function in machine learning. This function gives an indication of the model's level of confidence in the correctness of its prediction for a certain input by assigning a numerical value to its output. This score is determined by a number of variables, such as the model's internal representations, the distribution of the training data, and how well it understands how closely the input resembles the training samples. The ability to define thresholds for accepting or rejecting predictions in crucial settings is made possible by this confidence score, which is crucial in decision-making processes. Additionally, it aids in evaluating the model's uncertainty, making sure that predicted probabilities are calibrated correctly, and directing ensemble approaches that integrate many models. The dependability and interpretability of the model's predictions in real-world applications are influenced by how different models and tasks calculate confidence scores.

2.4. SILENT FAILURE

The term "silent failure" in image classification describes scenarios in which an algorithm or system incorrectly classifies an image without producing any obvious signs of its mistake. To put it another way, the failure happens without any obvious mistakes, alarms, or cautions that would make a human reviewer doubt the veracity of the algorithm's output.

When it comes to medical imaging, where precise diagnoses and decisions are essential for patient treatment, this phenomenon is very alarming. Silent failures can result in inaccurate diagnosis, bad treatment strategies, and jeopardised patient safety. Healthcare professionals may intuitively believe the algorithm's output because there aren't any obvious signs of failure, which makes it challenging for them to recognise when an error has happened.

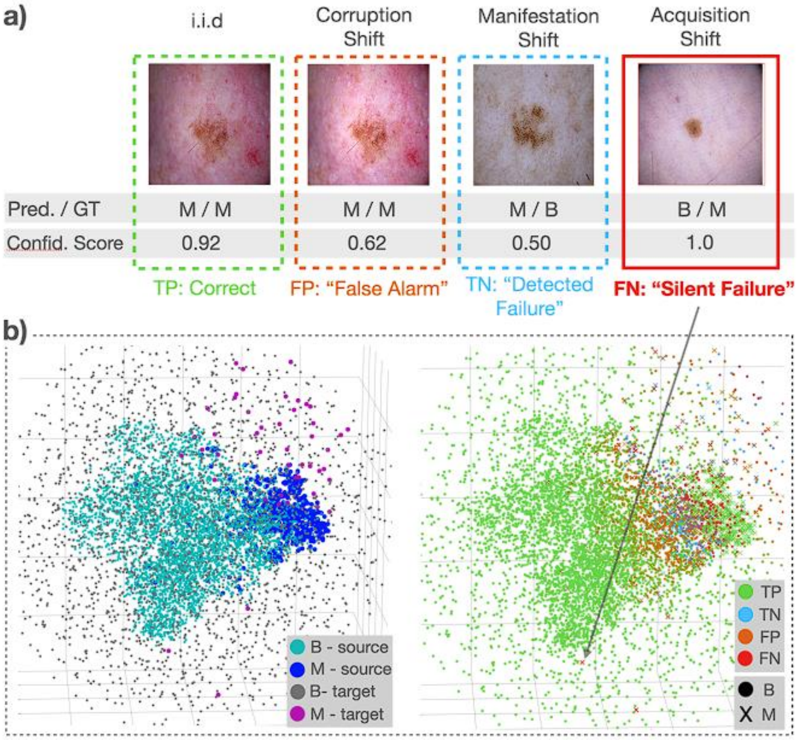


Figure 1: Silent Failure (T. Bungert et al., 2023).

To maintain the dependability and efficiency of automated systems in healthcare, it is crucial to identify and correct hidden faults in medical image classification. This requires the creation of reliable validation procedures, ongoing observation, and the incorporation of fail-safe systems.

A mix of innovative deep learning methods, rigorous testing, and a profound comprehension of the clinical environment are required to address silent failures. The objective is to develop algorithms that are highly accurate as well as transparent, comprehensible, and equipped with systems to sound alarms when uncertainties or potential errors appear. The study of medical image categorization can help make medical procedures safer and more dependable by reducing the incidence of silent failures.

2.5. Deep Learning Classification Model

A variety of neural network architectures are used in various deep learning classification models to automatically learn and recognise patterns, features, and representations from data in order to execute image classification tasks. The study of computer vision has substantially improved because to these models. We'll explore a few well-known architectures.

2.5.1. AlexNet

A cutting-edge deep convolutional neural network (CNN) architecture created specifically for image categorization applications is called AlexNet. It rose to fame by winning in the 2012 ImageNet Large Scale Visual Recognition Challenge. AlexNet, developed by Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton, has three fully linked layers at the bottom of five convolutional layers. The use of rectified linear units (ReLU) for activation, data augmentation to minimise overfitting, dropout regularisation, and the use of GPUs for quicker training are notable innovations that were made. With its advanced design, these developments, and a sizable dataset, AlexNet was essential in launching the deep learning era and laying the foundation for later CNN systems.

2.5.2. VGG

The VGG design has a focus on both deep and basic convolutional neural networks. The construction of VGG models is consistent, with numerous layers of 3x3 convolutional filters followed by layers that pool data to a maximum. Although VGG networks need a large number of parameters, their depth helps them catch fine-grained data.

2.5.3. MobileNet

MobileNet models are created to operate effectively on portable and embedded technology. They use depth-wise separable convolutions to increase accuracy while lowering computing costs. There are several variations of MobileNet architectures, including MobileNetV2 and MobileNetV3, each with improved performance and efficiency.

2.5.4. EfficientNet

A set of models called EfficientNet tries to strike a balance between accuracy and processing economy. These models use a compound scaling strategy to scale the network's depth, width, and resolution, producing models that are both effective and efficient for a variety of tasks.

2.5.5. ResNet-50

A deep neural network called ResNet-50 has 50 layers, including fully connected, convolutional, batch normalisation, activation, and activation layers. The network can learn residual features thanks to the inclusion of residual connections, also known as skip connections. The vanishing gradient issue is lessened by these connections, enabling the training of extremely deep networks.

ResNet-50 has achieved outstanding performance on numerous image classification benchmarks and excels in capturing complex image features.

3. Proposed Method

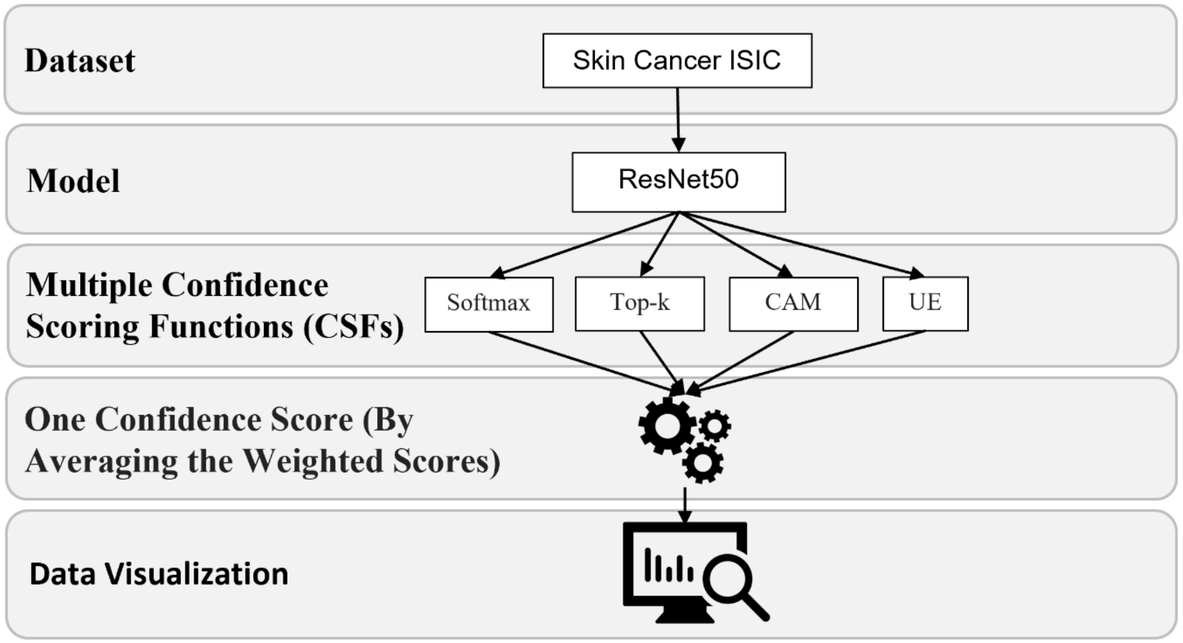

Dataset
Skin Cancer ISIC
Model
ResNet50
Multiple Confidence Scoring Functions (CSFs)
Softmax
Top-k
CAM
UE
One Confidence Score (By Averaging the Weighted Scores)
Data Visualization

Figure 2: ResNet50 Model with Multiple Confidence Score Function

The suggested architecture is a sophisticated ensemble model that integrates four different confidence scoring methods to increase the accuracy of predictions provided by a ResNet50 model. The goal is to develop a more reliable and precise classification system.

The ResNet50 backbone, a pre-trained convolutional neural network used for its capacity to extract significant characteristics from images, forms the basis of the design. This fundamental component transforms the input images into intermediate features that contain various visual attributes.

Four confidence scoring functions are combined, each presenting a different viewpoint on the classification assignment, and this integration results a new viewpoint. An accepted technique for calculating class probabilities is provided by the "Softmax Probability Scores". Indicating how firmly the model distinguishes between potential classes, the "Top-k Probability Margin" takes the difference between the top-k projected class probabilities into account. By highlighting the areas of an image that are most important for a class prediction, "Class Activation Maps (CAM) Score" promotes a greater comprehension of the model's decision-making. Finally, "Uncertainty Estimation" allows for a more detailed evaluation of reliability by quantifying uncertainty and capturing the model's confidence in its predictions.

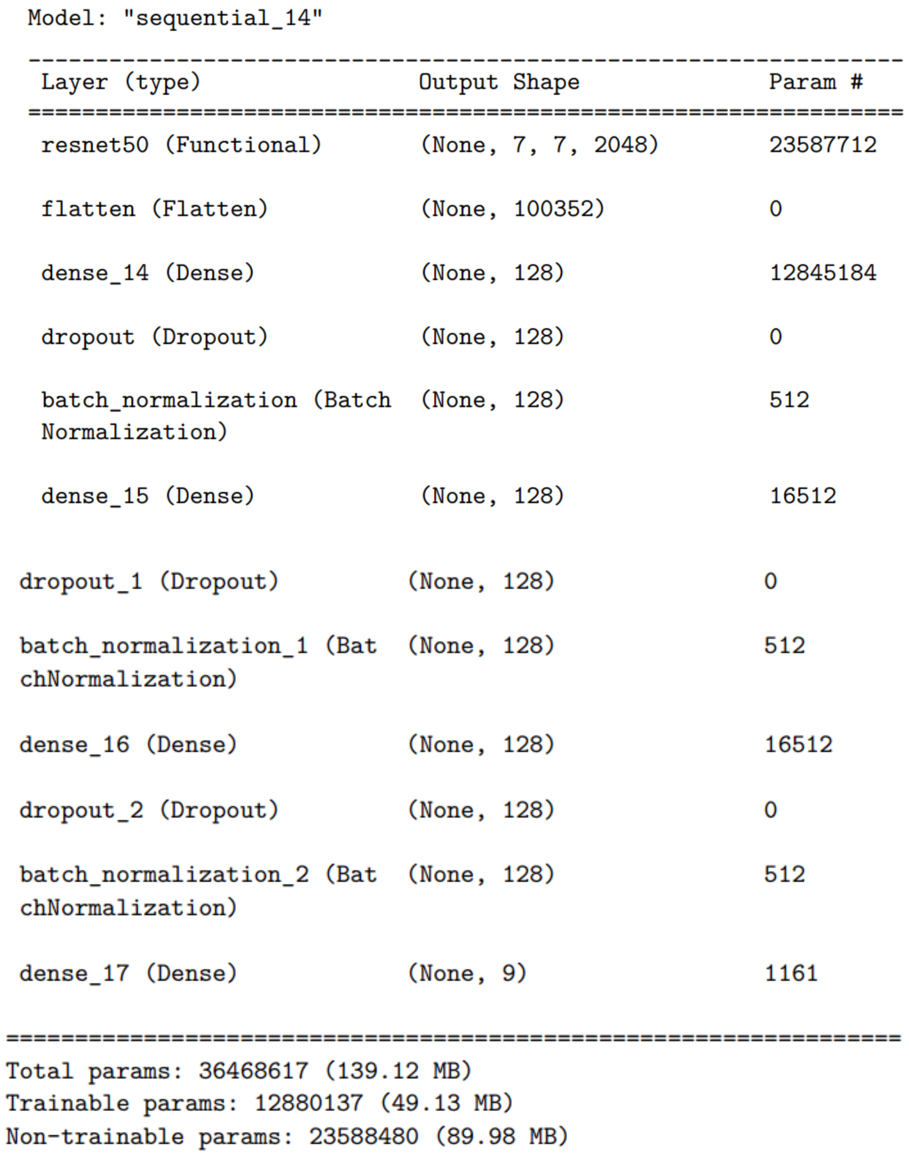
Following the generation of these various confidence ratings, the architecture concentrates on combining them into a single, comprehensive "Combined Confidence Score." This is done by using a weighted total, where each confidence score's normalised and scaled value is multiplied by a weight to represent its relative relevance. The unified combined score is then created by averaging the weighted scores.

This architecture has numerous advantages. The model makes predictions by incorporating various confidence score functions and drawing on a variety of data. This strengthens the model's resistance to different problems, resulting in more precise and assured classifications. By enabling practitioners to precisely control the impact of each scoring function, the weighted sum technique improves the effectiveness of the entire decision-making procedure. Consequently, this ensemble model can handle complex scenarios, varying data, and ambiguous inputs better.

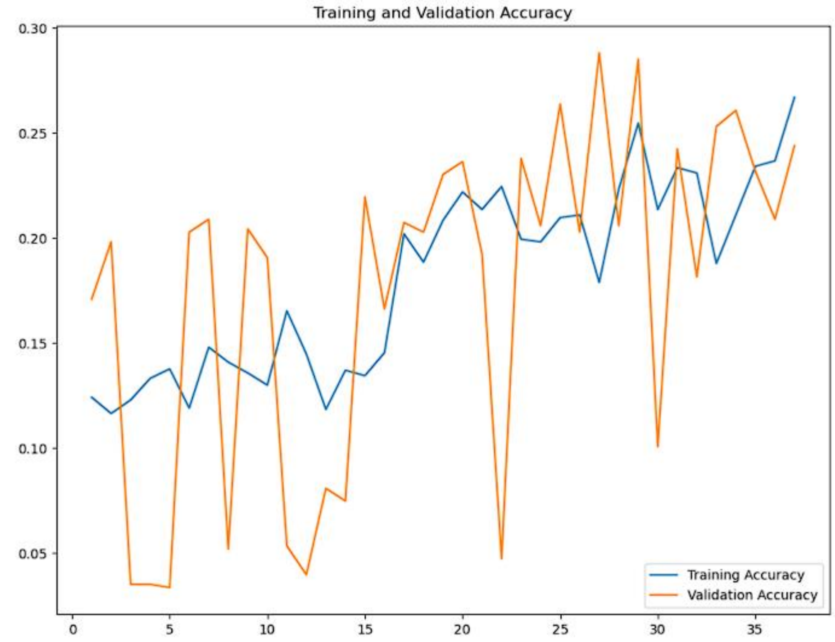
However, due diligence is necessary. The selection of weights for each function is crucial and necessitates repeated testing on validation datasets. To ensure the ensemble's effectiveness, training and validation require a large amount of labelled data. To confirm that the combined scoring system does, in fact, improve the model's predictive ability, ongoing evaluation and validation are crucial.

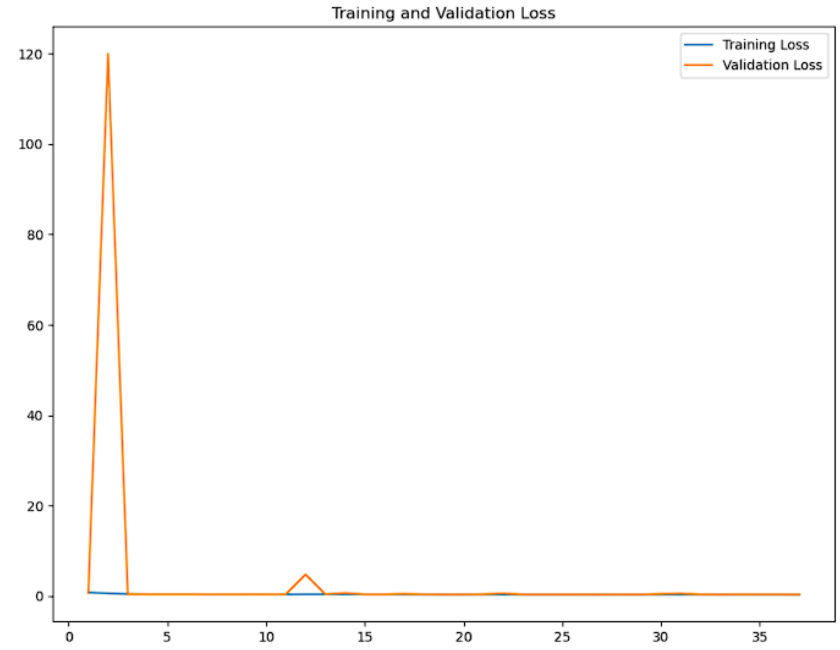
A ResNet50 model's capabilities can be improved, allowing for more informed and secure predictions in multi-class image classification scenarios. This architecture ultimately demonstrates the potential of merging multiple perspectives through different confidence scoring functions.

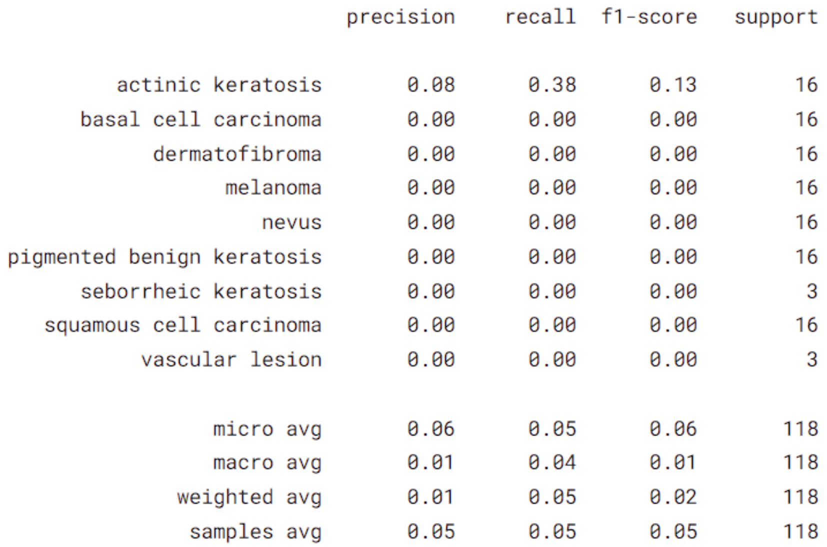
4. Results



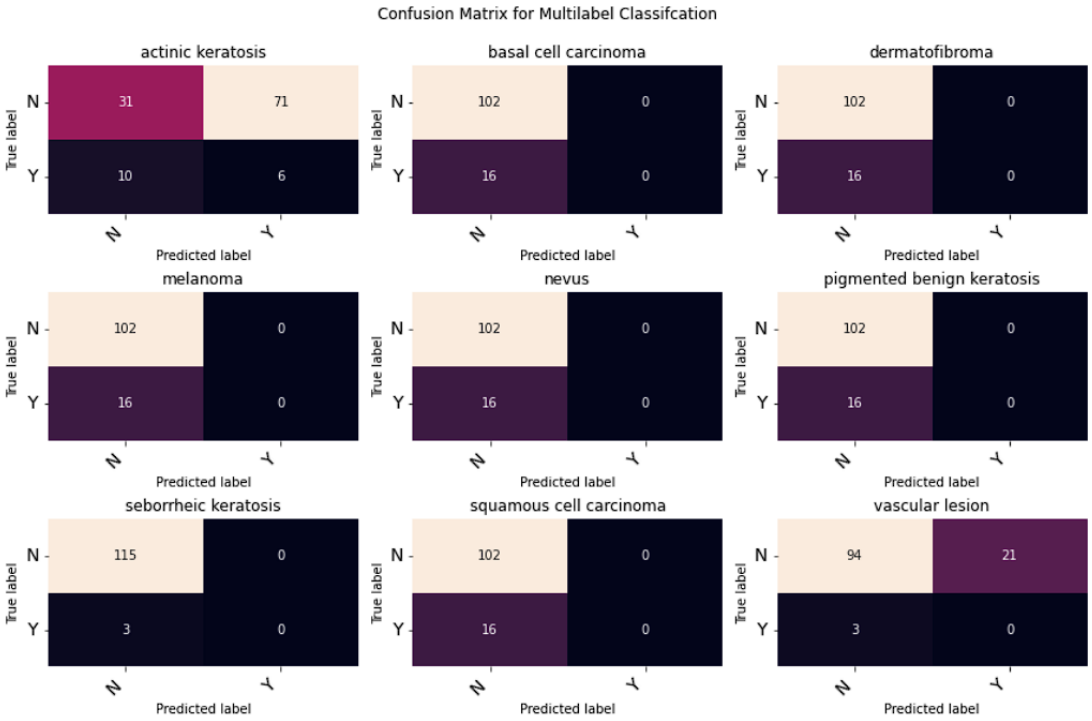
```
Model: "sequential_14"

 _________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 resnet50 (Functional)       (None, 7, 7, 2048)        23587712

 flatten (Flatten)           (None, 100352)            0

 dense_14 (Dense)            (None, 128)               12845184

 dropout (Dropout)           (None, 128)               0

 batch_normalization (Batch  (None, 128)               512
 Normalization)

 dense_15 (Dense)            (None, 128)               16512

 dropout_1 (Dropout)         (None, 128)               0

 batch_normalization_1 (Bat  (None, 128)               512
 chNormalization)

 dense_16 (Dense)            (None, 128)               16512

 dropout_2 (Dropout)         (None, 128)               0

 batch_normalization_2 (Bat  (None, 128)               512
 chNormalization)

 dense_17 (Dense)            (None, 9)                 1161

=================================================================
Total params: 36468617 (139.12 MB)
Trainable params: 12880137 (49.13 MB)
Non-trainable params: 23588480 (89.98 MB)
```


Training and Validation Accuracy
0.30
0.25
0.20
0.15
0.10
0.05
0
5
10
15
20
25
30
35
Training Accuracy
Validation Accuracy


Training and Validation Loss
120
100
80
60
40
20
0
0
5
10
15
20
25
30
35
Training Loss
Validation Loss



```
                          precision    recall  f1-score   support

       actinic keratosis       0.08      0.38      0.13        16
    basal cell carcinoma       0.00      0.00      0.00        16
          dermatofibroma       0.00      0.00      0.00        16
                melanoma       0.00      0.00      0.00        16
                   nevus       0.00      0.00      0.00        16
pigmented benign keratosis       0.00      0.00      0.00        16
    seborrheic keratosis       0.00      0.00      0.00         3
 squamous cell carcinoma       0.00      0.00      0.00        16
         vascular lesion       0.00      0.00      0.00         3

               micro avg       0.06      0.05      0.06       118
               macro avg       0.01      0.04      0.01       118
            weighted avg       0.01      0.05      0.02       118
             samples avg       0.05      0.05      0.05       118
```



In this study, the efficacy of a ResNet50 model in the difficult task of classifying skin cancer lesions is thoroughly evaluated. The model was carefully trained and thoroughly evaluated utilising a broad dataset that included nine different kinds of skin lesions in order to do this. A fundamental grasp of the model's fundamental prediction skills was provided via the evaluation process, which included the traditional assessment metrics of loss and accuracy.

To improve the performance of the model, the study went further by using modern confidence scoring methods. These included methods that intended to offer a more accurate and refined estimation of the model's predictions, such as softmax Probability Scores, Top-k Probability Margin, Class Activation Maps (CAM) Score, and Uncertainty Estimation. Notably, in order to accurately represent the model's overall confidence in its predictions, these individual confidence values were carefully combined into a single composite score by a weighted sum.

The ResNet50 model has an accuracy level of nearly 0.2054, according to the outcomes of the model evaluation phase, and a loss of about 0.3877. Following that, a probability matrix with 118 rows (representing distinct images) and 9 columns (each designating a distinct class) was created using the model's predictions. When these predictions were contrasted with the ground truth labels produced from the test image file locations, a startling contrast was seen. The ground truth label was given a value of 1 in this process, while all other labels were given a value of 0. This resulted in the formulation of a matrix.

The creation of a thorough multilabel classification report provided more insight into the model's performance. Heatmaps were used to visually represent the generated confusion matrix, providing a clear illustration of the model's advantages and disadvantages across multiple classes. The classification report also provided specific metrics for each class, including precision, recall, F1-score, and support values. Unfortunately, our in-depth examination showed that the majority of classes had poor performance indicators.

5. DISCUSSION

The analysis of the ResNet50 model's performance offers a glimpse into the difficulties, complications, and possible restrictions that characterise the field of medical image analysis.

The moderate accuracy of the ResNet50 model on the various skin lesion dataset is the initial focus. The challenging job of accurately classifying skin lesions, which frequently exhibit subtle differentiations, is highlighted by the accuracy of about 20.54% that was obtained. Due to the potential diagnostic relevance of minor changes in lesion appearance, dermatological classification is inherently complex. This highlights the seriousness of the medical setting, where accuracy is crucial, and the requirement for strong and trustworthy classification models.

One noteworthy aspect of the study is its attempt to improve predictions through the use of sophisticated confidence score methods. A proactive approach to improving the model's predictions may be seen in the integration of methods includes softmax Probability Scores, Top-k Probability Margin, Class Activation Maps (CAM) Score, and Uncertainty Estimation. However, the slight increases in accuracy that followed their adoption produced an important realisation. Even though these methods aim to capture subtle differences, it's possible that they don't fully take into account how difficult it is to classify skin lesions. Depending on how well the model's architecture captures the underlying intricacies, their performance impact may vary.

The composite score, which is calculated from the weighted sum of confidence scores, is at the heart of the discussion. The results of the study show that even this careful aggregation did not result in appreciable improvements in accuracy. This demonstrates the delicate balancing act between a model's complexity and its ability to generalise across different kinds of skin lesions. It emphasises how important it is to fine-tune the hyperparameters and align the scoring function with the unique features of the dataset.

This study emphasises how data science and clinical knowledge can work together in the broader field of medical image analysis. Incorporating machine learning techniques with the observations and insights of domain experts is essential. These specialists are in a unique position to recognise crucial visual clues that may enable more precise classification of skin lesion groups. The study also emphasises the need for comprehensive methods that cover feature extraction, collecting data, and model architecture.

6. CONCLUSION

This work is a testament to the complexity of skin cancer lesion classification and the difficulties inherent in medical image processing. While demonstrating the ResNet50 model's promise, its performance also highlights how difficult it is to convert computational power into precise medical diagnoses. Although useful, the incorporation of higher confidence scoring systems highlights the need for a deeper comprehension of how they interact with underlying data complexities.

The results of the study require an extensive review of model development. The collaboration between clinical practitioners and data scientists is essential when machine learning is integrated into healthcare. The industry stands to acquire more thorough insights into accurate skin cancer classification by combining domain experience with cutting-edge algorithms.

This work adds to the notion that medical image analysis is a multifaceted endeavor that integrates technology with human skill rather than merely being a technical goal. In the field of healthcare, it highlights the ongoing requirement for constant study, collaboration, and innovation to promote accurate diagnosis and better patient outcomes.

7. References

Calisto, F. M., Santiago, C., Nunes, N., & Nascimento, J. C. (2022). BreastScreening-AI: Evaluating medical intelligent agents for human-AI interactions. *Artificial Intelligence in Medicine*, *127*, 102285.

Calisto, F. M., Santiago, C., Nunes, N., & Nascimento, J. C. (2021). Introduction of human-centric AI assistant to aid radiologists for multimodal breast image classification. *International Journal of Human-Computer Studies*, *150*, 102607.

Shamshirband, S., Fathi, M., Dehzangi, A., Chronopoulos, A. T., & Alinejad-Rokny, H. (2021). A review on deep learning approaches in healthcare systems: Taxonomies, challenges, and open issues. *Journal of Biomedical Informatics*, *113*, 103627.

Tschandl, P., Rinner, C., Apalla, Z., Argenziano, G., Codella, N., Halpern, A., ... & Kittler, H. (2020). Human–computer collaboration for skin cancer recognition. *Nature Medicine*, *26*(8), 1229-1234.

Bernhardt, M., Ribeiro, F. D. S., & Glocker, B. (2022). Failure detection in medical image classification: A reality check and benchmarking testbed. *arXiv preprint arXiv:2205.14094*.

Bungert, T. J., Kobelke, L., & Jaeger, P. F. (2023). Understanding Silent Failures in Medical Image Classification. *arXiv preprint arXiv:2307.14729*.

8. Contribution

	MD. Sazib Ahmed	**Masud Pervez**	**Badhan Akter**	**Contribution (%)**
	20-42076-1	*20-42656-1*	*20-42225-1*	
Conceptualization	55%	25%	20%	100 %
Data curation	100%	0%	0%	100 %
Formal analysis	30%	30%	40%	100 %
Investigation	40%	30%	30%	100 %
Methodology	50%	25%	25%	100 %
Implementation	50%	20%	30%	100 %
Validation	30%	40%	30%	100 %
Theoretical derivations	30%	30%	40%	100 %
Preparation of figures	30%	40%	30%	100 %
Writing – original draft	40%	20%	40%	100 %
Writing – review & editing	50%	30%	20%	100 %