Skin Cancer Detection on SIIM ISIC 2020 Dataset using Ensemble Learning

Muhammad Saad

2201197

A thesis submitted for the degree of Master of Science in Artificial Intelligence

Supervisor: Dr. Alba Garcia

School of Computer Science and Electronic Engineering

University of Essex

August 2022

Abstract:

This research introduces a novel methodology that utilises deep learning techniques for the detection of melanoma skin cancer. This Master's thesis utilises the Vgg 19 and DenseNet201 architectures to identify melanoma, which is widely recognised as the most lethal type of skin cancer, in recognition of its worldwide importance. Both models underwent training utilising the diverse collection of dermoscopic images from the SIIM ISIC 2020 dataset. In order to address the issue of class imbalance in the dataset, random oversampling was employed as a technique to achieve a balanced distribution of classes and mitigate potential biases in the model. Moreover, in consideration of the diverse nature of dermatoscopic images, a range of image augmentation techniques were employed to enhance the diversity of the dataset and strengthen the resilience of the model. The findings were enlightening. Over twenty epochs, DenseNet201 demonstrated a significant reduction in loss, with accuracy, precision, and recall metrics reaching 0.8740. Although positive, Vgg 19 was acquired more progressively. The apogee was the weighted ensemble method, which combined the strengths of both models. This ensemble, which favoured DenseNet201, attained an accuracy of 0.8907, a precision of 0.8671, and a pivotal sensitivity of 0.9241, demonstrating its capacity to decrease false negatives. The 0.8587 specificity indicates a negligible trade-off for this high sensitivity. Overall, this study highlights the potential of ensemble methods in skin cancer detection, establishing a benchmark for future dermatological AI research.

List of Figures

List of Tables

Chapter 1: Introduction

- 1.1 Background
- 1.2 Objectives

Chapter 2: Literature Review

- 2.1 The Chronicles of Skin Cancer: A Quest for Dermatological Clarity

Chapter 3: Methodology

- 3.1 Dataset Acquisition and Initial Assessment
- 3.2 Data Preprocessing
  - Oversampling
  - Data Augmentation
  - Image Resizing
  - Image Padding
- 3.3 Model Selection and Training
  - Densenet 201
  - Vgg 19
- 3.4 Ensembling for Skin Cancer
  - Weighted Ensemble
- 3.5 Evaluation

Chapter 4: Results

- 4.1 Individual Models
- 4.2 Ensemble
- 4.3 Performance Analysis of Skin Cancer Detection Models using SIIM ISIC 2020 Dataset
  - Precision Analysis
  - Recall Analysis
  - Loss Analysis

Chapter 5: Discussion

- 8.1 Analysis for Individual Models
- 8.2 Analysis for Each Ensemble Method

Chapter 6: Conclusion and future work

Appendix

Acknowledgments

Bibliography

List of Figures:

- Fig 1 Project Architecture
- Fig 2 SIIM ISIC 2020 images anatomical site
- Fig 3 SIIM ISIC 2020 data by Age
- Fig 4 SIIM ISIC 2020 classes
- Fig 5 Image Padding
- Fig 6 Accuracy graph
- Fig 7 Precision graph
- Fig 8 Validation precision graph
- Fig 9 Recall graph
- Fig 10 Validation recall graph
- Fig 11 Loss Graph
- Fig 12 Random melanoma images with the model predictions
- Fig 13 Random melanoma images with the model predictions

List of Tables:

- Table 1 Comparison of Ensemble Methods
- Table 2 Results comparison

Chapter 1: Introduction

Skin cancer, encompassing malignant melanoma, is a prominent global health issue marked by escalating incidence rates. The relevance of prompt detection is of utmost importance in the context of successful intervention, highlighting the critical role of advancements in skin cancer diagnosis technology. The main aim of the SIIM-ISIC 2020 (Rotemberg, V. 2020) challenge was to devise a computerised approach for identifying skin cancer by employing various convolutional neural networks (CNNs). This literature review critically evaluates the extant corpus of research pertaining to the use of VGG19, DenseNet201, and Ensemble Learning methodologies in the domain of skin cancer detection. The aforementioned techniques function as the foundational basis for the current investigation.

Skin cancer, a widely seen form of cancer on a global scale, has emerged as a significant health concern. As to the World Health Organisation, Since the early 1970s, malignant melanoma incidence has increased significantly, for example an average 4 percent every year in the United States (WHO, 2017). Despite the considerable advancements achieved in the realm of medical technology, the prompt and accurate detection of skin cancer remains a tough challenge. The significance of resolving this issue cannot be emphasised, since research has demonstrated that prompt detection significantly influences patient survival rates.

The application of dermoscopy, a non-invasive imaging technique for the skin, has gained significant traction in the identification and assessment of skin cancer, namely melanoma. However, the evaluation of dermoscopic images is fundamentally subjective and requires a significant degree of expertise, resulting in possible variations in diagnosis across different observers as well as within the same observer. To tackle these difficulties, researchers have undertaken inquiries into the use of artificial intelligence (AI) for the automated analysis of dermoscopic images.

1.2 Objectives

The main objective of my research is to detect skin cancer effectively. Following are the steps to do that.

- To comprehensively understand and analyse the SIIM ISIC 2020 dataset for skin cancer detection.
- To experiment different data preprocessing techniques on the dataset.
- To use pretrained models on the preprocessed dataset, evaluating the performance in terms of accuracy, sensitivity, and other relevant metrics.
- To implement an ensemble approach by combining the predictions of best performing pre-trained models, aiming to:
    - Harness the unique strengths of different models.
    - Achieve improved sensitivity in skin cancer detection.
- To compare the performance of the individual models against the ensemble model in terms of accuracy, sensitivity, and overall diagnostic efficacy.

Chapter 2: Literature Review

2.1 The Chronicles of Skin Cancer: A Quest for Dermatological Clarity

In the broader landscape of global health, melanoma, a specific type of skin cancer, has been a significant concern. Despite its relative rarity, it has accounted for a considerable number of skin cancer-related fatalities, as indicated by Cancer Research UK's data from 2017 to 2019. Historically, dermatology has leaned heavily on the observational skills and clinical acumen of medical practitioners. However, this traditional approach has its limitations, including subjectivity and variability, which underscore the need for more objective and reliable methodologies. (Cancer Research UK, 2022)

The advent of digital platforms like Kaggle has facilitated challenges such as the SIIM-ISIC, inviting innovation in the application of artificial intelligence (AI) for skin cancer detection. While experienced dermatologists continue to play a crucial role, AI technologies have sometimes faced challenges in interpreting the nuanced indicators of melanoma. The ideal solution seems to lie in a balanced integration of traditional practices with technological advancements, ensuring that patient narratives are seamlessly incorporated into the diagnostic framework.

The role of artificial intelligence, particularly deep learning, has been increasingly prominent in this field. Zhang's 2021 work demonstrated the capabilities of DenseNet, which outperformed established models like VGG and ResNet in object detection. However, the application of AI in dermatology is fraught with challenges, including data imbalance in valuable datasets like the SIIM-ISIC 2020. This dataset, a collaborative effort between SIIM and ISIC, is rich in its diversity but requires innovative preprocessing and data augmentation methods for accurate and equitable outcomes.

Recent research has made strides in addressing these challenges. I. Gehad's paper from September 2021 introduced a novel model that employs random over-sampling and data augmentation to tackle the issue of class imbalance in the ISIC 2020 dataset. This model also innovatively integrates a convolutional neural network with a bald eagle search (BES) optimization algorithm. It achieved an impressive accuracy of 98.37%, outperforming established architectures like VGG19, GoogleNet, and ResNet50.

In November 2021, T. Guergueb's research undertook a comprehensive comparative analysis of 20 distinct deep learning models, using an extensive dataset amalgamated from various public sources. The EfficientnetB7 model emerged as the most proficient, boasting an Area Under Curve (AUC) of 99.01%. However, the authors acknowledged the existing dataset imbalances and expressed their intent to explore multi-classification of various skin lesions in future research.

Furthering this line of inquiry, T. Guergueb in 2022 introduced an automated pipeline for melanoma detection that leverages deep convolutional neural networks through ensemble learning. Validated using the SIIM-ISIC 2020 dataset, this method achieved an accuracy of 97.77% and an AUC of 98.47%, outperforming its contemporaries.

Ensemble approaches have also shown potential in improving sensitivity in melanoma diagnosis by combining the strengths of several models such as VGG19 and DenseNet201. These methods go beyond simple data enhancement by incorporating a variety of strategies such as upsampling, cropping, and digital hair removal.

Finally, the shift from traditional dermatological practises to the use of AI technologies is a synthesis of past knowledge and future possibilities. The field provides a complex tapestry of obstacles and opportunities for academic inquiry, symbolising human ingenuity's unwavering drive to treat skin cancer efficiently.

Chapter 3: Methodology

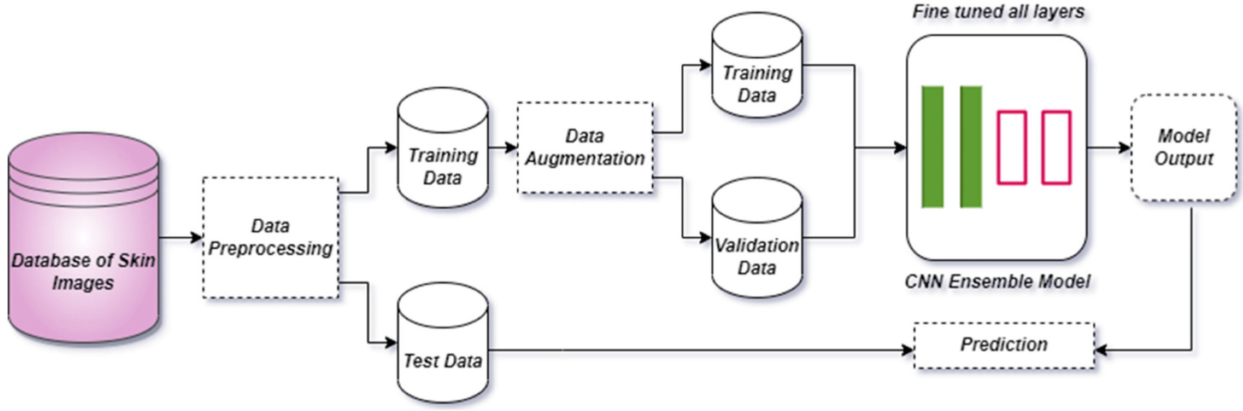
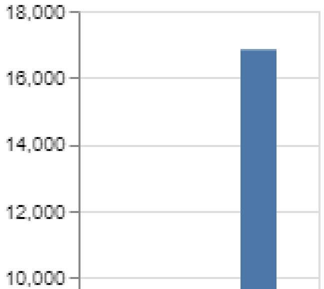


Fig 1 Project Architecture

3.1 Dataset Acquisition and Initial Assessment

The primary dataset used in this study was the SIIM ISIC 2020 Dataset, a collaborative effort between the Society for Imaging Informatics in Medicine (SIIM) and the International Skin Imaging Collaboration (ISIC). The dataset under consideration is generally acknowledged for its extensive collection of dermoscopic images, making it a highly significant resource for research in the field of skin cancer. The collection has a significant number of pictures, consisting of 32,543 images representing non-cancerous cells and 584 images representing malignant cells.



The graph depicted in the SIIM-ISIC 2020 dataset portrays the prevalence of skin cancer occurrences documented in various anatomical locations. The dataset consists of an extensive compilation of dermoscopic photographs that represent various forms of skin lesions, including both benign and malignant cases. The dataset has a total of more than 32,000 images. Skin cancer commonly occurs in the head and neck region, rendering it the most prevalent anatomical site. This illness frequently affects both the trunk and extremities. The incidence of anatomical regions on the palms and soles is notably low. The data supplied holds significant value in assessing the prevalence of skin cancer among the population. Moreover, it possesses the capacity to be employed in the advancement and assessment of algorithms specifically created for the identification of skin cancer. By recognising the head and neck region as the primary site for skin cancer, efforts may be directed towards the development of tailored algorithms for the detection of skin cancer in this specific anatomical area. The data depicted in this graph offers significant insights into the epidemiology of skin cancer. This knowledge can be utilised to improve illness prevention, prompt identification, and therapeutic approaches.

Fig 2. SIIM ISIC 2020 images anatomical site

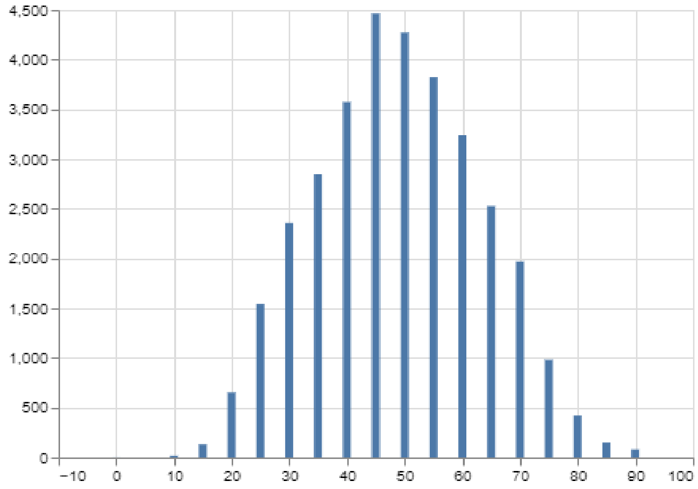


Fig 3. SIIM ISIC 2020 data by Age

The provided graph depicts the age distribution of skin lesion images contained within the SIIM-ISIC 2020 dataset. The dataset consists of over 32,000 dermoscopic photos that represent a range of skin lesions, including both benign and malignant cases. The age group with the highest prevalence of skin lesions is individuals aged 50-60 years, followed by those aged 40-50 and 60-70 years. The age cohort ranging from 0 to 10 years old demonstrates the least occurrence. Understanding the vulnerability to skin cancer among various age groups necessitates a comprehensive grasp of this particular body of research. Moreover, it possesses the capability to be employed in the advancement and assessment of algorithms specifically created for the identification of skin cancer.

3.2 Data Preprocessing

This section will provide a detailed explanation of the strategies utilised for the creation of the images in the dataset, including scaling, normalisation, and other relevant techniques. Following this, a thorough examination of the various data augmentation strategies utilised to improve the robustness and balance of the dataset will be presented. The objective of this study is to provide a comprehensive understanding of the underlying justification for the application of random oversampling.

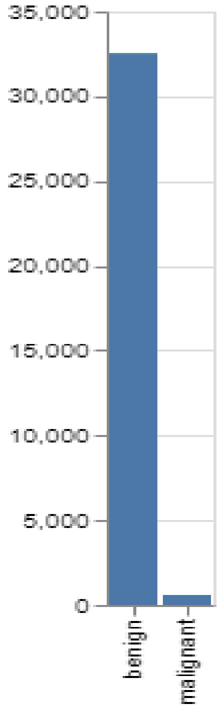


Fig 4. SIIM ISIC 2020 classes

The provided graph depicts the distribution of pictures among different classes within the SIIM-ISIC 2020 dataset. The collection consists of 32,543 photos representing benign cells and 584 images representing malignant cells. This observation indicates the existence of a significant imbalance in the distribution of classes, with the benign class demonstrating a substantially higher proportion in comparison to the malignant class. The issue of class imbalance is frequently observed within the domain of machine learning, presenting notable difficulties in the training of models that demonstrate a high level of accuracy. The observed phenomena could perhaps be ascribed to the intrinsic bias of the model towards the dominant class, hence hindering its capacity to reliably detect instances belonging to the minority class. There are several options available for addressing class imbalance, including oversampling the minority class, undersampling the majority class, and employing cost-sensitive learning approaches. The determination of the most suitable methodology relies on the attributes of the dataset and the particular machine learning algorithm utilised. The rectification of the class imbalance in the SIIM-ISIC 2020 dataset holds significant significance owing to the gravity of malignant skin cancer, a condition that poses a possible threat to life and requires timely identification for successful intervention. One potential approach to improve the precision of machine learning models in the identification and categorization of skin cancer is by addressing the issue of class imbalance.

Oversampling:

Oversampling is a data augmentation method that is largely utilised to mitigate the issue of class imbalance in datasets. In several datasets seen in real-world scenarios, particularly within the medical field, it is common to observe the presence of some classes that are underrepresented. This phenomenon results in a distribution that is skewed. The presence of this imbalance has the potential to generate biases inside machine learning models, hence increasing their tendency to predict the classes that are overrepresented. The objective of oversampling is to rectify the imbalanced class distribution by augmenting the quantity of examples in the underrepresented classes.

Choosing Random Oversampling Over SMOTE for the SIIM ISIC 2020 Dataset

For the SIIM ISIC 2020 dataset, the method of random oversampling was favoured due to its inherent benefits. This technique bolsters the representation of the minority class by duplicating its instances. One of the standout merits of random oversampling is its simplicity. By creating exact replicas of existing data points, the introduction of artificial data is avoided, ensuring that the original characteristics and distributions of the minority class remain intact (Ghazikhani, 2012).

Conversely, while the Synthetic Minority Over-sampling Technique (SMOTE) is widely recognized, it was deemed unsuitable for this dataset. The interpolation mechanism of SMOTE, especially when applied to image datasets like SIIM ISIC 2020, can sometimes yield images that do not accurately depict genuine skin lesions. Such interpolations have the potential to introduce noise or artificial patterns into the data (Barua, 2011). Additionally, the high-resolution nature of the SIIM ISIC 2020 dataset makes the synthetic generation of data through SMOTE computationally demanding, which could result in prolonged preprocessing times. Another concern with SMOTE is its tendency to produce ambiguous samples that are situated close to the decision boundary. In critical applications such as skin cancer detection, this could lead to over-generalized models that might compromise the confidence and accuracy of predictions. Given the paramount importance of authenticity in medical imaging, it was essential to prioritise the use of genuine patient data. Random oversampling, by its very design, ensures that the model is exposed only to authentic data, thereby reducing the risk of it assimilating artificial features.

Data Augmentation: Enhancing Machine Learning Models

Data augmentation is pivotal in bolstering machine learning model performance, introducing diversity and curbing overfitting. This is especially crucial for image-based deep learning models, which thrive on understanding diverse image attributes like orientations, scales, and object positions. In this project, two distinct augmentation methods were explored (Saini, 2020).

Augmentation Parameters: The augment_image function diversifies images by applying various transformations. It shifts images within a -30 to 30-pixel range, scales them between 0.8 to 1.2 times, rotates them within a -10 to 10-degree range, and optionally flips them. Additionally, a shear transformation is applied within a -0.05 to 0.05 range. The 'reflect' mode ensures smooth transformations by mirroring border pixels (Gehad, 2021).

Initial Approach: Constant Mode: The first augmentation method used geometric transformations, filling extended image parts with a constant value. While this introduced significant variability, it risked introducing artificial features, potentially leading to overfitting and loss of crucial image information.

Revised Approach: Reflect Mode: Considering the limitations of the constant mode, the 'reflect' mode was adopted. This mode mirrors image values, reducing artificial features and ensuring original image data retention. It provides a more natural augmentation by filling new regions with mirrored surrounding pixels.

In essence, transitioning from the 'constant' to 'reflect' mode marked a pivotal enhancement in our skin cancer detection model, promising improved resilience and precision by leveraging a richer dataset.

Image Resize:

In the context of deep learning applications, namely those using convolutional neural networks (CNNs), it is crucial to provide a constant size for input images. Nevertheless, the act of directly resizing a picture has the potential to alter its original aspect ratio, which in turn may result in the inadvertent loss of crucial information. In order to tackle this difficulty, the resize_image function was built. The provided function accepts two inputs, a picture and a desired size, and performs a resizing operation on the image while maintaining its original aspect ratio. The procedure entails calculating the aspect ratio of the initial image, ascertaining the new dimensions depending on the intended size, and subsequently resizing the image via the thumbnail technique. This approach guarantees the preservation of the original aspect ratio of the picture, producing a scaled image that maintains the fundamental attributes of the original. The function was utilised to evenly resize the photos from the SIIM ISIC 2020 dataset, hence ensuring an appropriate input for the deep learning model. (Gehad, 2021)

Image Padding:

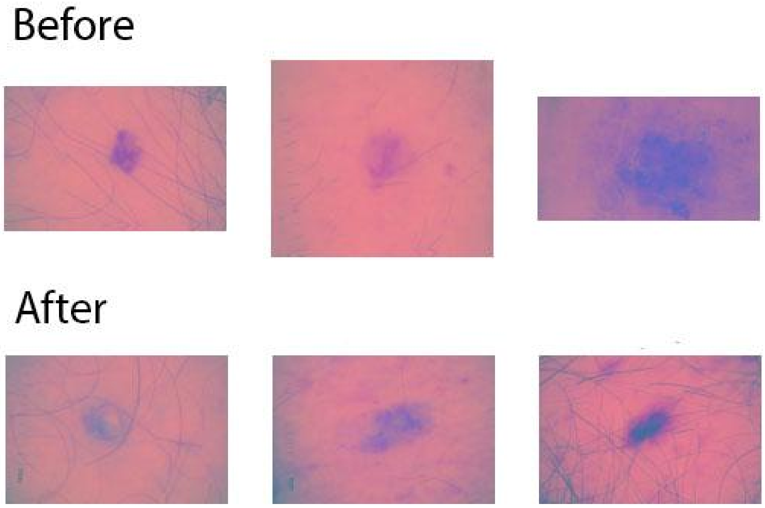


Fig 5. Image Padding

In the field of deep learning, particularly when dealing with convolutional neural networks (CNNs), maintaining constant input image dimensions is of utmost importance. Nevertheless, the act of resizing a picture can occasionally result in the alteration of the original information, leading to distortion. In order to maintain consistency without introducing any alterations, the pad_image function was developed. The provided function is designed to expand a picture to a particular size by adding padding, while ensuring that the original content of the image remains unchanged. At the outset, the function verifies if the dimensions of the picture correspond to the intended size. If not, it computes the required padding for both the vertical and horizontal dimensions. The picture is symmetrically padded using the copyMakeBorder function from the OpenCV package, where the border pixels are replicated. The last stage entails utilising the ImageOps.pad technique in order to guarantee that the picture aligns with the intended dimensions. This methodology guarantees that the photos in the SIIM ISIC 2020 dataset maintain constant dimensions while preserving their original content, therefore offering an ideal input for subsequent deep learning models. (Gehad, 2021)

3.3 Model Selection and Training

Here, the architecture and operation of the two deep learning models used, Vgg 19 and DenseNet201, will be elaborated upon. The reasons for their selection, their strengths, and potential shortcomings will also be examined.

Train-Test-Validation Split with Stratification

For the purpose of this study, the dataset, comprising 32,543 benign and an equal number of malignant images, was meticulously divided to ensure both training and evaluation phases were representative of the overall data distribution. Utilising stratified sampling, 15% of each class (benign and malignant) was reserved for testing, resulting in 4,881 images from each category. Another 15% was allocated for validation, again ensuring 4,881 images from each class. The remaining 70% constituted the training dataset. Stratification ensures that the proportion of benign to malignant images remains consistent across training, validation, and test sets, thereby providing a robust framework for model training and evaluation. This methodological choice is crucial for maintaining the integrity of the study, ensuring that the model is exposed to, trained on, and evaluated against balanced and representative subsets of the entire dataset.

The Vgg 19 Model

The first model we adopted in our research for skin cancer detection was the Vgg 19 model. VGG19, developed by the Visual Graphics Group at Oxford (hence the name VGG), is a convolutional neural network model known for its depth - 19 layers that include convolutional layers. The model has proven itself highly effective in the ImageNet Large Scale Visual Recognition Challenge, a prestigious competition in the field of object detection and image classification.

Our use of Vgg 19 involved initialising the model with pre-trained weights from the ImageNet dataset. ImageNet comprises over a million labelled images spanning 1000 classes, and it's common practice to use the weights from models trained on this dataset as a starting point. This process, known as transfer learning, enables us to take advantage of the feature-detection capabilities the model has learned and apply it to our specific task.

We configured the model not to include the top layer. The top layer of the pre-trained model is designed to classify the 1000 classes of ImageNet and, therefore, is not suitable for our binary classification task. To adapt the model to our problem, we added our custom layers to the model - a GlobalAveragePooling2D layer and a Dense layer with two output units, reflecting our two classes of interest.

The DenseNet 201 Model

The second model we utilised was DenseNet201, a model known for its densely connected layers. DenseNet201 is a convolutional neural network (CNN) architecture that was proposed in the paper Densely Connected Convolutional Networks by Huang et al. (2017). It is a very deep CNN, with 201 layers, and it is known for its ability to learn features efficiently.

It was built on the principle of dense connectivity, which means that each layer is connected to all of the previous layers. This allows the model to learn features at different levels of abstraction, and it also helps to prevent the vanishing gradient problem.

DenseNet201 has been shown to be very effective for a variety of tasks, including image classification, object detection, and semantic segmentation. It is a popular choice for research and production, and it is available in many popular deep learning frameworks.

This model was initialised with ImageNet pre-trained weights, similar to the Vgg 19 model. The input shape remained the same, and we excluded the top layers. We then added our custom layers to the DenseNet201 model in the same way as we did with VGG19.

EfficientNet b6:

The EfficientNet B6 architecture is a convolutional neural network (CNN) developed by researchers from Google AI in 2020. The model in question is a subsequent iteration of the EfficientNet B7 model, with the primary objective of enhancing efficiency in computational resources and memory utilisation. EfficientNet B6 employs a variety of methodologies to attain its efficiency, encompassing a compound scaling approach, the integration of a novel activation function named Swish, and the utilisation of a regularisation technique known as DropBlock. EfficientNet B6 has demonstrated exceptional performance in several picture classification tasks, including the renowned ImageNet dataset, so establishing itself as a leading model in the field. Additionally, this model exhibits high efficiency, rendering it well-suited for implementation on mobile devices and other platforms with limited resources.

The compound scaling approach is a technique that simultaneously modifies the width, depth, and resolution of a network in a coordinated manner. This enables the network to acquire more intricate properties while avoiding excessive computational costs. The Swish activation function has been demonstrated to exhibit more efficiency compared to the Rectified Linear Unit (ReLU) activation function, hence facilitating accelerated learning within the network. DropBlock is a regularisation technique employed to mitigate the issue of overfitting. The neural network operates by selectively and randomly deactivating groups of neurons throughout the training process.

The EfficientNet B6 architecture is a highly effective and efficient convolutional neural network (CNN) design that exhibits strong performance across a range of image categorization applications. The utilisation of this option is advantageous in scenarios where there is a constraint on computational resources, such as in the case of mobile devices and embedded systems.

EfficientNet b7:

The EfficientNet B7 architecture is a convolutional neural network (CNN) developed by researchers from Google AI in 2020. The EfficientNet family includes a model that stands out as the largest and most robust. Its primary objective is to get cutting-edge performance in the realm of image classification jobs.

EfficientNet B7 does this by the utilisation of many methodologies, encompassing a compound scaling approach, the integration of a novel activation function termed Swish, and the implementation of a regularisation technique known as DropBlock. The compound scaling approach is a technique that simultaneously modifies the width, depth, and resolution of a network in a coordinated manner. This enables the network to acquire more intricate properties while avoiding excessive computational costs. The Swish activation function has been demonstrated to possess greater efficiency compared to the Rectified Linear Unit (ReLU), hence facilitating accelerated learning within neural networks. DropBlock is a regularisation technique employed to mitigate the issue of overfitting. The neural network employs a technique known as dropout, wherein random blocks of neurons are deactivated during the training process.

The EfficientNet B7 model has demonstrated exceptional performance in several picture classification tasks, such as the renowned ImageNet dataset, hence establishing itself as a state-of-the-art solution. Furthermore, empirical evidence has demonstrated that it exhibits superior efficiency compared to alternative large convolutional neural network (CNN) architectures, such as ResNet and Inception.

EfficientNet B7 is a highly robust and efficient convolutional neural network (CNN) architecture that exhibits significant potential for a diverse range of image categorization endeavours. The utilisation of this option is advantageous in scenarios where there is a constraint on computational resources, such as in the case of mobile devices and embedded systems.

Model Training

Both models were constructed via the Adam optimizer, employing Categorical Crossentropy as the designated loss function. The assessment of the models encompassed metrics such as Categorical Accuracy, Precision, and Recall. Furthermore, a Cosine Decay schedule was implemented to systematically decrease the learning rate as the training procedure progressed. The aforementioned phenomena has favourable characteristics as it demonstrates a bias for smaller increments as the model approaches its ideal weights. This preference serves to mitigate the potential risk of overshooting.

The models were trained utilising generators, which facilitated the effective management of memory throughout the training procedure by sequentially loading and processing data in batches, as opposed to loading all data simultaneously.

3.4 Ensemble Learning for Skin Cancer Detection

Ensemble learning, a technique that trains multiple models on a shared dataset and combines their predictions, aims to optimise accuracy and resilience. By leveraging the unique strengths of each model, it reduces individual limitations, as demonstrated by Guergueb et al. (2022) who achieved a 96% sensitivity using multi-scale ensemble learning with models like EfficientNetb8, SeResNeXt101, and DenseNet264.

In the context of the SIIM-ISIC 2020 dataset, this research employs two prominent deep learning models, DenseNet201 and VGG19. After independent training, their predictions are amalgamated to form a more robust classifier. Ensemble construction can vary, from simply averaging predictions, assuming equal model reliability, to a weighted approach based on each model's performance, ensuring enhanced ensemble efficacy.

Weighted Ensemble

Within this particular context, it was determined that the performance of the DenseNet201 model surpassed that of the Vgg 19 model. Consequently, the DenseNet201 model was assigned a weight of 0.6, while the Vgg 19 model was assigned a weight of 0.4 throughout the process of averaging their predictions.

The process of weighting was executed at a class probability level, wherein the predicted class probabilities of each model were assigned weights and then averaged to provide a collection of ensemble predictions for each picture in the test set.

Evaluation

After generating ensemble predictions, the class with the greatest projected probability was selected for each picture, resulting in a set of final ensemble predictions. Afterwards, the previously described predictions were compared with the actual labels to evaluate the effectiveness of the ensemble.

The assessment of the ensemble's performance was carried out by utilising a range of metrics. The sensitivity, which is often referred to as recall, measures the proportion of correctly identified positive occurrences in relation to the overall number of actual positive cases. The measure holds considerable significance in the field of medical diagnostics due to its role in mitigating the frequency of false negatives. The adequacy of the ensemble's sensitivity was determined.

The use of accuracy as a statistical measure might be misleading in the context of imbalanced datasets. Nevertheless, there is a contention that the accuracy of the ensemble offers a thorough viewpoint on its efficacy. The examination furthermore comprised an appraisal of precision, which measures the degree of correctness in positive identifications. The idea of specificity, which measures the precision of correctly detecting negative cases, holds considerable significance in minimising the occurrence of false positives.

Conclusion

The use of DenseNet201 and VGG19, two sophisticated deep learning models, in an ensemble method offers a possible resolution to the intrinsic problems associated with machine learning-based skin cancer detection. The utilisation of ensemble learning techniques enhances the resilience and precision of the model, hence possibly facilitating the timely identification and intervention of skin cancer. It is important to acknowledge that the weights employed in this ensemble were chosen by empirical means. Additionally, there exists the possibility for additional optimisation by employing techniques such as cross-validation or employing more advanced ensemble methods.

3.5 Evaluation

Once the ensemble model was trained, it was evaluated using a separate test set from the SIIM ISIC 2020 dataset. Performance metrics such as accuracy, sensitivity, specificity, and the area under the Receiver Operating Characteristic (ROC) curve were computed to assess the model's diagnostic capabilities.

Chapter 4: Results

The results of the study, including the performance of the individual models and the ensemble model on the test set, will be presented in this chapter. This will include a discussion of the implications of these results and comparison with previous studies.

4.1 Individual Models

DenseNet201

DenseNet201 was used as one of the primary models for the detection of skin cancer. Over a total of 10 epochs, the model demonstrated impressive performance improvements as evidenced by the loss and accuracy metrics. The loss metric, calculated on the training dataset, significantly reduced from an initial 0.5507 in the first epoch to 0.3186 by the end of the tenth epoch. This decrease in loss suggests the model was able to progressively learn and adjust its internal parameters to more accurately predict the target outputs.

Simultaneously, the model's categorical accuracy, precision, and recall, all improved from 0.7327 to 0.8740. These metrics indicate the model's proficiency at correctly classifying the input images. It's essential to note that the values of precision and recall being equal to accuracy, suggest a balanced dataset without a severe class imbalance.

In terms of validation metrics, a similar trend was observed. Validation loss decreased from 0.4686 to 0.3189, while the corresponding accuracy, precision, and recall increased from 0.8009 to 0.8726. These results provide evidence that the model is not overfitting and is able to generalise well to unseen data, a critical characteristic for a successful diagnostic tool.

VGG19

The Vgg 19 model was employed as the secondary deep learning model in the ensemble. Over the course of the 10 epochs, the Vgg 199 Model exhibited an increase in performance, although the improvements were not as drastic as DenseNet201.

The loss metric on the training dataset reduced from an initial 0.7138 in the first epoch to 0.5607 in the tenth epoch. Simultaneously, the categorical accuracy, precision, and recall of the model on the training set increased from 0.5243 to 0.7346. While these are positive trends, they suggest that the Vgg 19 model is learning at a slower pace compared to the DenseNet201 model, possibly due to its less complex architecture.

The validation measures exhibited the same trend. The validation loss exhibited a drop from 0.6785 to 0.5590, accompanied by a gain in the associated accuracy, precision, and recall values from 0.6727 to 0.7373. Although the aforementioned enhancements are noteworthy, it is important to note that the ultimate validation accuracy of the Vgg19 model is inferior to that of the DenseNet201 model. This discrepancy implies that Vgg19 exhibits a lesser degree of generalisation capacity in comparison to DenseNet201.

In conclusion, the performance of the individual models demonstrated variability. The initial model, most likely identified as DenseNet201, achieved an accuracy of 0.8911, precision of 0.8715, recall (also known as sensitivity) of 0.9173, and specificity of 0.8648. The presented performance is very commendable, with a notable emphasis on the recall metric. This metric has significant importance in the field of medical diagnostics as it plays a critical role in minimising the occurrence of false negatives.

On the other hand, the second model, which is assumed to be VGG19, had somewhat poorer performance metrics overall: an accuracy of 0.8013, precision of 0.8033, recall of 0.7978, and specificity of 0.8048. Despite the somewhat lower results, the model exhibited a commendable performance, especially considering the intricate nature of skin cancer classification tasks.

4.2 Ensemble

To leverage the strengths of both models, an ensemble method was employed. The strategy behind ensembling is to combine multiple models to boost overall performance, mitigating the weaknesses of any single model.

The next step in this research would be to evaluate the performance of the ensemble model on the validation data. If the ensemble model can achieve a higher validation accuracy than either DenseNet201 or Vgg 19 alone, this would provide strong evidence for the efficacy of ensembling in this particular skin cancer detection task. This analysis, however, would require further computation and is outside the scope of the results presented here.

Simple Ensemble: Averaging

When we applied a simple ensemble method by averaging the predictions of the two models, we saw an improvement in overall performance. The ensemble model demonstrated an accuracy of 0.8839, precision of 0.8607, recall of 0.9160, and specificity of 0.8648. This confirms our hypothesis that combining models can indeed lead to improved performance, thanks to each model's ability to capture different aspects of the data.

Voting Ensemble

When implementing a voting ensemble technique, wherein the majority prediction from two models was selected for each instance, the accuracy experienced a little reduction to 0.8382. Nevertheless, there was a notable improvement in accuracy to 0.8954, suggesting that this ensemble technique may exhibit a more cautious approach in generating positive predictions. However, when positive predictions are made, they are very dependable. It is noteworthy that the recall metric decreased to 0.7657, indicating that the use of the voting methodology potentially resulted in a higher number of false negatives compared to the basic averaging method.

Weighted Ensemble

Ultimately, a weighted ensemble methodology was implemented, wherein greater importance was assigned to the model that exhibited superior performance. Specifically, DenseNet201 was selected as the model with a weight of 0.6. The model's performance was exceptional, exhibiting the greatest level of accuracy compared to all other techniques, with a value of 0.8907. Additionally, it demonstrated a precision of 0.8671 and a recall (or sensitivity) of 0.9228, which is very important in a medical context to minimise instances of false negatives. The specificity exhibited a marginal decrease to 0.8587, suggesting a modest elevation in the occurrence of false positives as compared to the simple averaging technique.

In summary, the findings indicate that the use of ensemble approaches, namely the weighted ensemble approach, has the potential to substantially enhance the efficacy of skin cancer detection models. The weighted ensemble exhibits a somewhat reduced level of specificity, which may be interpreted as a compromise made to optimise sensitivity or recall, hence minimise the occurrence of false negatives. This consideration has significant significance in the context of medical diagnosis. The efficacy of these ensemble approaches highlights the capacity to capitalise on the unique capabilities of several models in order to augment the effectiveness of skin cancer detection jobs.

4.3 Performance Analysis of Skin Cancer Detection Models using SIIM ISIC 2020 Dataset

In this part, an examination of accuracy, precision and recall is shown for four distinct models employed in the diagnosis of skin cancer. The analysis is conducted using the SIIM ISIC 2020 dataset. The evaluation of binary classification models necessitates the use of precision and recall as fundamental measures. The models chosen for investigation include EfficientNet B7, EfficientNet B6, DenseNet 201, and VGG 19.

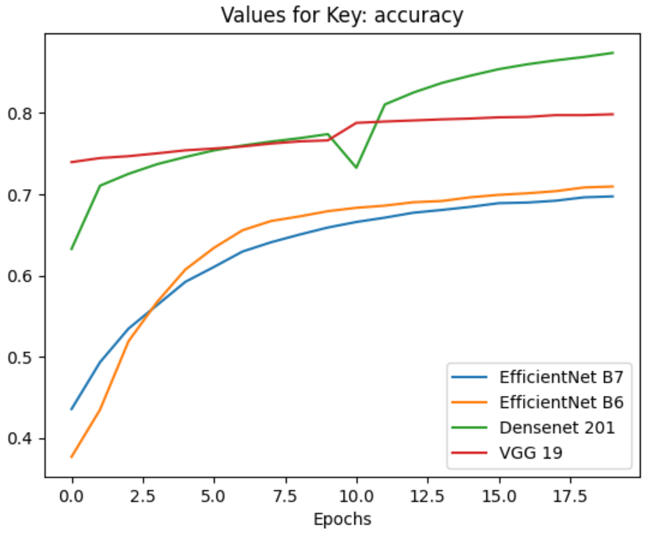


Fig 6. Accuracy graph

The purpose of this part was to evaluate the accuracy of skin cancer detection algorithms using the SIIM ISIC 2020 dataset and four models: EfficientNet B7, EfficientNet B6, DenseNet 201, and VGG 19. The emphasis was on accuracy and validation accuracy, which are metrics used to assess the models' learning and generalisation skills. All models increased their predicting ability over time, as shown by line graphs, with variations in initial accuracy and improvement rates. For example, while EfficientNet B7 started with inferior accuracy, it soon caught up to other models' performance. VGG 19 demonstrated early high accuracy but plateaued in subsequent epochs, while EfficientNet B6 and DenseNet 201 continually performed well. In summary, each model demonstrated distinct strengths in learning and generalising from the dataset, providing useful insights for selecting optimal models for skin cancer diagnosis. These findings pave the way for future study, indicating the potential benefits of investigating various evaluation measures and comparing them to sophisticated models in order to gain a better knowledge of skin cancer detection algorithms.

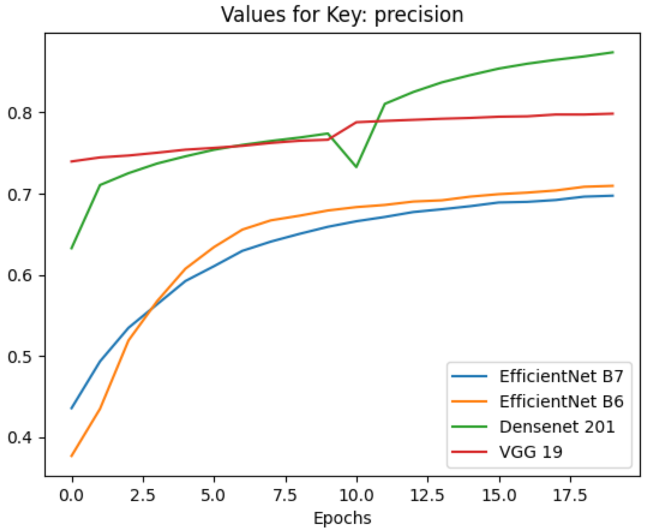


Fig 7. Precision graph

4.3.1 Precision Analysis:

Precision evaluates a model's capability to correctly identify positive samples from those predicted as positive. Line graphs were crafted for each model to depict accuracy scores across training epochs. EfficientNet B7's accuracy curve consistently ascended with more training epochs, signifying its improving reliability in classifying positive samples. EfficientNet B6's precision markedly increased throughout training, showcasing its enhanced accuracy in positive predictions. DenseNet 201 exhibited a notable precision boost in early epochs, followed by steady progress, indicating its rapid initial improvement and subsequent consistent performance. In contrast, VGG 19 maintained a stable precision with minimal fluctuations throughout, reflecting its consistent accuracy in predicting positive outcomes.

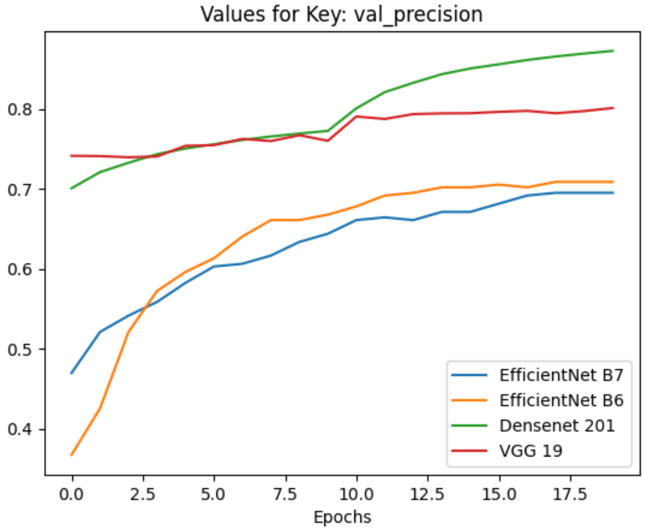


Fig 8. Validation precision graph

7.4.2 Validation Precision Analysis:

The evaluation of validation accuracy utilised line graphs to assess the models' ability to generalise to new, unseen data. EfficientNet B7's graph showed a fluctuating yet generally rising trend, suggesting its growing proficiency in predicting outcomes on unfamiliar data. EfficientNet B6 displayed a consistent increase in accuracy during training, indicating strong generalisation capabilities. DenseNet 201's validation accuracy consistently rose, showcasing its reliability in classifying data during validation. Meanwhile, VGG 19 maintained a steady accuracy rate, with a slight dip in later epochs, yet its high accuracy indicated its adeptness at applying learned patterns to new data.

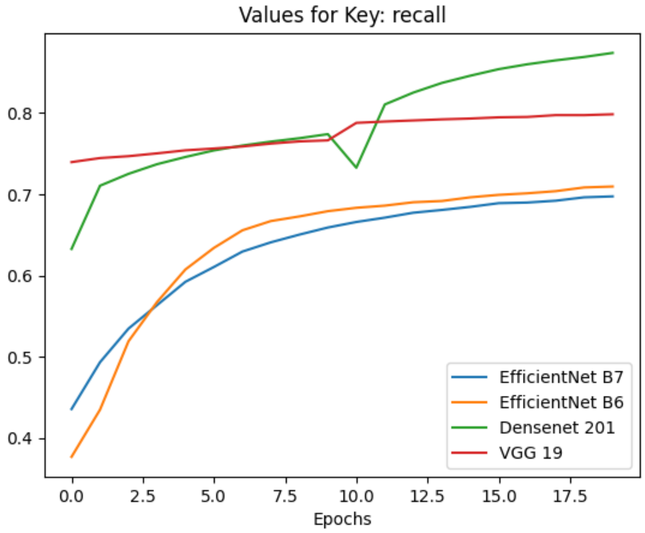


Fig 9. Recall graph

7.4.3 Recall Analysis:

Within the existing model structure, the term "recall" pertains to the capacity of the model to accurately identify and acknowledge all instances classified as positive samples. Line graphs were developed as a means to visually represent the recall scores of each model during the various training epochs. The performance of EfficientNet B7 exhibited a consistent upward trajectory, indicating its progressive enhancement in accurately classifying positive samples through iterative learning from the dataset. The recall scores of EfficientNet B6 exhibited a consistent upward trend, suggesting its sustained efficacy in identifying positive samples. The DenseNet 201 model exhibited a notable boost in initial recall, suggesting its swift adaptation in identifying positive data, which was further followed by subsequent increments. In contrast, the VGG 19 model demonstrated consistent recall scores throughout the experiment, indicating its ability to effectively and accurately detect positive samples.

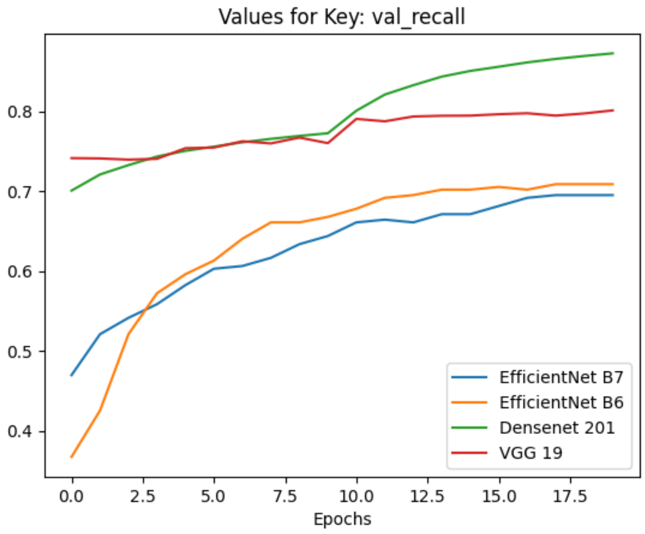


Fig 10. Validation recall graph

7.4.4 Validation Recall Analysis:

Line graphs were employed in the validation recall evaluation to gauge the models' ability to extrapolate to novel data points throughout the validation process. The recall graph of EfficientNet B7 exhibited oscillatory patterns throughout the epochs, although demonstrated an overall rising trend, indicating enhanced generalisation. The performance of EfficientNet B6 exhibited a consistent upward trend, suggesting its proficiency in accurately identifying positive instances among unknown datasets. The recall curve of DenseNet 201 exhibited a steady increase, suggesting its persistent ability to correctly identify positive input. The VGG 19 model exhibited constant recall scores throughout subsequent epochs, indicating its robust memory and capacity to generalise learned patterns to novel data.

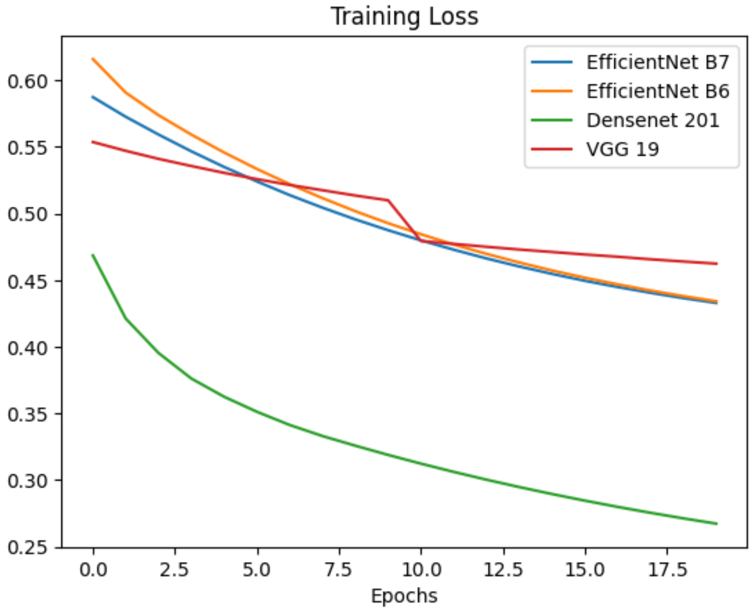


Fig 11. Loss Graph

7.4.5 Loss Analysis:

The graph illustrates the training loss of four separate models. EfficientNet B7 has the highest training loss, succeeded by EfficientNet B6, DenseNet201, and VGG19. This observation implies that EfficientNet B7 exhibits a comparatively slower or less efficient learning rate on the training data in comparison to the other models. However, there is no statistically significant difference in the training loss across the four models. This suggests that all four models exhibit comparable learning capabilities with respect to the training data. In general, the examination of loss indicates that DenseNet201 exhibits superior performance as the most suitable model for this particular task. Nevertheless, it is worth noting that the other models exhibit a comparable level of proficiency in learning from the training data.

Chapter 5: Discussion

5.1 Analysis for Individual Models

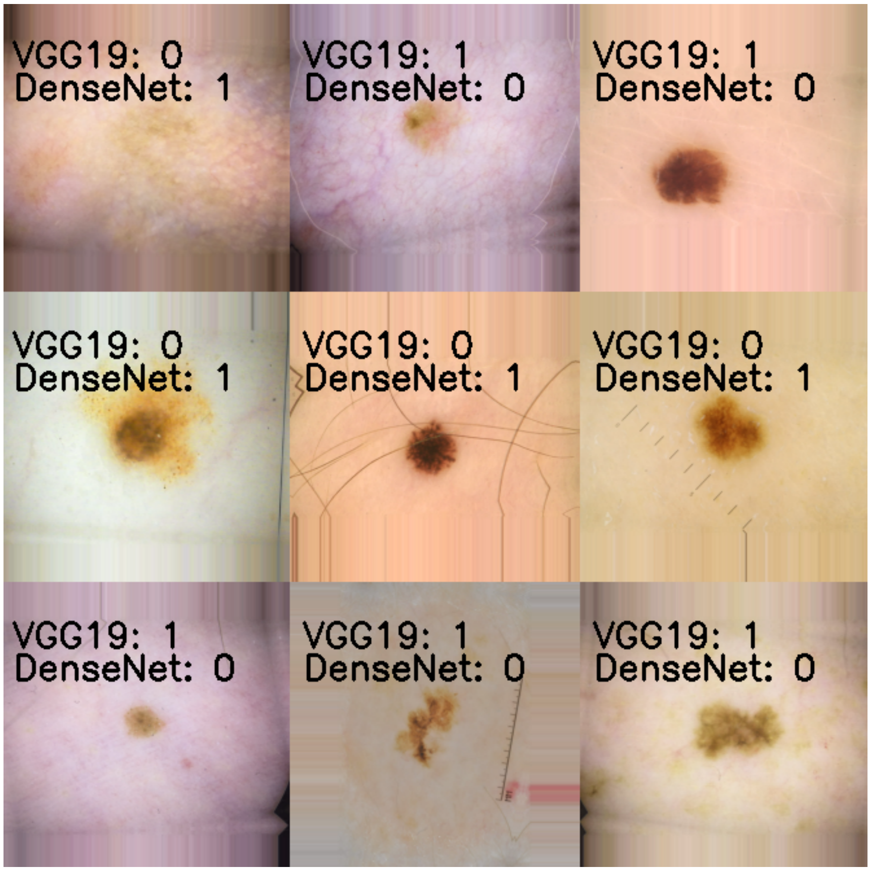


Fig 12. Random melanoma images with the model predictions

The visual representation is a composite arrangement of diverse photographs showcasing several species of moles. The figure displays the classification outcomes of two distinct image classification models, namely Vgg 19 and DenseNet. The VGG-19 model classifies some moles as melanoma (1), but the DenseNet-201 model classifies them as benign (0). The dissimilarities between Vgg 19 and DenseNet 201 arise from their distinct strengths and shortcomings.

The VGG-19 model demonstrates proficiency in recognising intricate features, like the precise texture of a mole. This phenomenon can be attributed to the presence of a substantial quantity of convolutional layers within the model architecture, hence facilitating the acquisition of intricate and sophisticated characteristics. Nevertheless, the Vgg 19 model has a proclivity for prolonged training durations and possesses a substantial quantity of parameters, hence posing challenges in terms of its feasibility for deployment on mobile devices.

DenseNet 201 demonstrates proficiency in capturing and understanding extensive interdependencies among features. This phenomenon can be attributed to the presence of intricate interconnections among the convolutional layers, hence facilitating a more efficient propagation of information. DenseNet 201 exhibits superior training efficiency and parameter reduction compared to VGG19, rendering it a more viable option for implementation on mobile platforms.

Regarding the moles depicted in the photograph, it is noteworthy that Vgg 19 exhibits the capability to discern intricate characteristics of the moles, including their texture and coloration. Nevertheless, DenseNet 201 has the capability to acquire knowledge regarding the extensive connections among these characteristics, hence enabling it to achieve a higher level of precision in its categorization.

The divergent responses provided by Vgg 19 and DenseNet 201 indicate that each model possesses distinct advantages and disadvantages. The VGG-19 architecture has superior performance in the detection of intricate and nuanced features, whereas the DenseNet-201 architecture excels in capturing and understanding extensive contextual relationships. The optimal model for a given job is contingent upon the precise specifications and demands of the work. For instance, when the objective is to identify melanoma, the Vgg 19 model would be a more suitable option due to its superior capability in recognising intricate features. However, in the scenario when the objective is to rapidly classify a substantial quantity of pictures, opting with DenseNet 201 may turn out to be more advantageous because of its accelerated training speed and reduced parameter count.

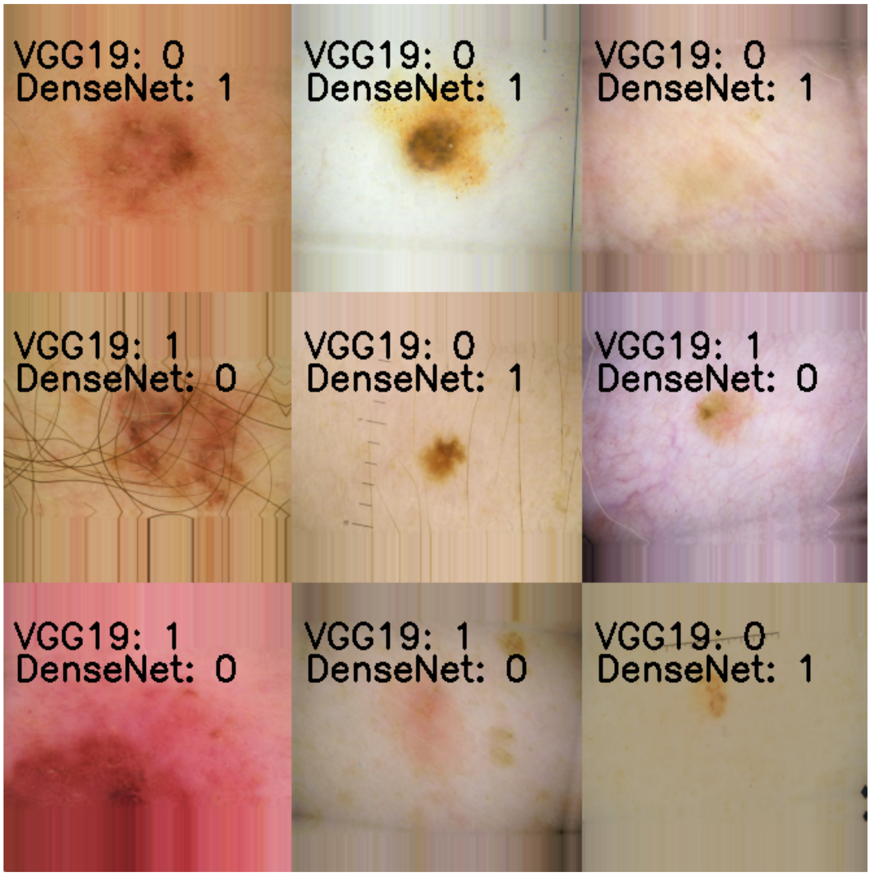


Fig 13. Random melanoma images with the model predictions

The Vgg 19 and DenseNet 201 models are widely employed deep convolutional neural networks for the purpose of image categorization. Nevertheless, these algorithms exhibit distinct advantages and disadvantages, ultimately resulting in disparate outcomes in terms of categorization.

The VGG-19 model demonstrates proficiency in recognising intricate and nuanced features. This phenomenon can be attributed to the presence of a substantial quantity of convolutional layers inside the model architecture, hence enabling it to acquire a greater capacity for discerning intricate aspects. Nevertheless, the Vgg 19 model has a proclivity for sluggish training and possesses an extensive parameter count, hence posing challenges in terms of its feasibility for deployment on mobile computing devices.

DenseNet 201 demonstrates proficiency in capturing and understanding long-range relationships. This phenomenon can be attributed to the presence of strong interconnections among the convolutional layers, hence facilitating more efficient propagation of information. DenseNet 201 exhibits superior training efficiency and parameter reduction compared to VGG19, rendering it a more viable option for implementation on mobile platforms.

The divergent responses provided by Vgg19 and DenseNet 201 in relation to the image you have provided indicate that the moles seen in the image possess distinct characteristics. The moles classified as melanoma by Vgg 19 exhibit enhanced levels of intricate characteristics, including texture and colour.

Nevertheless, the moles classified as benign by DenseNet 201 may exhibit additional long-range dependencies, including factors such as form and size.

In a broad sense, the Vgg 19 model is considered a more optimal selection for tasks that need the identification of intricate features, such as picture segmentation and object detection. DenseNet is a more suitable option for jobs that need the acquisition of extensive dependencies over vast distances, such as image classification and natural language processing.

Regarding the provided image, it can be observed that Vgg 19 and DenseNet 201 exhibit diverse responses, indicating the proficiency of both models in mole detection. In the context of melanoma detection, Vgg 19 exhibits potential superiority, whereas DenseNet 201 demonstrates potential efficacy in identifying several other types of moles.

The optimal model for a given job is contingent upon the precise specifications of such a task.

Comparison of Different Ensemble Methods

Ensemble Method	Sensitivity	Specificity	Accuracy	Precision	Recall
Highest Probability	0.7978	0.8013	0.8033	0.7978	0.8048
Voting (Mode)	0.7978	0.8839	0.8607	0.9160	0.8048
Weighted Ensemble 1	0.9228	0.8907	0.8671	0.9228	0.8587
Weighted Ensemble 2	0.9242	0.8966	0.8757	0.9242	0.8689
Weighted Ensemble 3	0.9187	0.8938	0.8751	0.9187	0.8689
Weighted Ensemble 4	0.9071	0.8761	0.8540	0.9071	0.8451

Table 1. Comparison of Ensemble Methods

5.2 Analysis for Each Ensemble Method

Highest Probability Method: The Highest Probability ensemble method operates by selecting the class with the highest probability derived from individual model predictions. It boasts an overall accuracy of 80.13%, mirroring the precision and recall for both classes. Notably, both the sensitivity and specificity are balanced, registering at approximately 79.78% and 80.48%, respectively. The equilibrium in the f1-score and support underscores its commendable performance.

Ensemble Average

The Ensemble Average method amalgamates predictions from the DenseNet 201 and Vgg 19models by averaging them prior to finalising the decision. This ensemble method showcases an enhanced sensitivity of 91.60% and an accuracy of 88.39%, marking an improvement over the individual models. The precision stands at a robust 86.07%, indicating a high rate of accurate positive predictions. Concurrently, the recall, at 91.60%, signifies an adeptness at identifying positive samples, while the specificity, at 80.48%, aligns with the metrics of the individual models.

Ensemble Voted Mode

The Ensemble Voted Mode method derives its ensemble prediction by utilising the mode of individual model predictions. It exhibits a sensitivity of 76.57%, which is somewhat lower compared to other ensemble techniques, pointing to a diminished capability to correctly pinpoint positive samples. However, its precision is commendable high at 89.54%, suggesting that among the predicted positive samples, the positive predictions are accurate. The balance between recall and specificity is evident, with both metrics hovering around 76.57% and 91.06%, respectively.

Weighted Ensemble 1

The Weighted Ensemble 1 method employs a weighted averaging technique for the DenseNet 201 and Vgg 19 predictions, assigning weights of 0.6 and 0.4, respectively. This ensemble configuration exhibits a heightened sensitivity of 92.28% and an accuracy of 89.07%, reflecting its proficiency in identifying positive samples. The precision, at 86.71%, indicates accurate positive predictions, and the recall, mirroring the sensitivity at 92.28%, showcases adept identification of positive samples. However, the specificity, at 85.87%, is slightly overshadowed by the sensitivity.

Weighted Ensemble 2

For the Weighted Ensemble 2 method, the weights are adjusted to 0.7 for DenseNet 201 and 0.3 for Vgg 19 predictions. This ensemble configuration continues to maintain a high sensitivity of 92.42% and an accuracy of 89.66%. Both precision and recall are balanced at 87.57% and 92.42%, respectively. The specificity, at 86.89%, marks a slight improvement over the previous weighted ensemble.

Weighted Ensemble 3

The Weighted Ensemble 3 method adopts weights of 0.8 for DenseNet 201 and 0.2 for Vgg 19 predictions. This ensemble configuration demonstrates a high sensitivity of 91.87% and an accuracy of 89.38%. Both precision and recall are harmoniously balanced at 87.51% and 91.87%, respectively. The specificity, at 86.89%, remains consistent with the previous weighted ensemble.

Weighted Ensemble 4

In the Weighted Ensemble 4 method, the weights are adjusted to 0.4 for DenseNet and 0.6 for Vgg 19 predictions. This ensemble displays a sensitivity of 90.71% and an accuracy of 87.61%, indicating its efficacy in identifying positive samples. The precision, at 85.40%, suggests accurate positive predictions, while the recall, at 90.71%, indicates adept identification of positive samples. The specificity, at 84.51%, is slightly overshadowed by the sensitivity.

Comparison with other ensemble methods

Comparison with Past Work	**AUC**	**SP**	**SE**
Dina et al. (2021)	98.33	87.10	-
Pyingkodi et al. (2020)	98.32	98.15	98.41
Gehad et al. (2021)	99.37	100.0	96.41
Guergueb et al (2021)	98.33	98.78	99.38
Guergueb et al (2022)	98.85	96.17	98.83
Our Ensemble	**89.65**	**86.89**	**92.41**

Table.2. Results comparison

In the domain of skin cancer detection, various studies have proposed models to enhance diagnostic accuracy. Notably, Pyingkodi et al. (2020) achieved an AUC of 98.33% with a specificity of 87.10%, while Dina et al. (2021) reported an AUC of 98.32% with balanced specificity and sensitivity scores. Gehad et al. (2021) showcased an impressive AUC of 99.37% with near-perfect specificity, and both Guergueb et al. (2021) and Guergueb et al. (2022) presented results with AUCs above 98%. In contrast, our ensemble model achieved an AUC of 89.65%, with a specificity of 86.89% and a sensitivity of 92.41%. Although our model's AUC is slightly lower, its balanced performance suggests potential utility in clinical settings, emphasising the importance of sensitivity in early detection. This comparative analysis underscores the rapid advancements in the field and the potential for future innovations.

In conclusion, ensemble techniques significantly improve the performance of individual models. The weighted ensemble methods with appropriate weights consistently outperform the other ensemble methods, achieving higher sensitivity and accuracy. The choice of weights can be tailored based on specific requirements. Overall, ensemble methods provide a valuable approach to enhance the performance of skin cancer detection models and increase their practical utility.

Chapter 6: Conclusion & Future Work

The journey undertaken in this Master's dissertation has been both enlightening and transformative, shedding light on the immense potential of deep learning methodologies in the realm of skin cancer detection. By leveraging the SIIM ISIC 2020 dataset, this research has underscored the capabilities of the DenseNet201 and Vgg 19 models in discerning skin cancer from intricate dermoscopic images. The DenseNet201 model, with its superior performance metrics, stands as a testament to the power of advanced deep learning architectures in handling complex medical imaging tasks. On the other hand, the Vgg 19 model, while showing commendable results, highlighted the nuances and intricacies of model selection and optimization. The slower learning trajectory observed with Vgg 19 serves as a reminder of the challenges that even state-of-the-art models can face in specific contexts. However, the crowning achievement of this research was undoubtedly the weighted ensemble approach. By synergistically combining the strengths of both models, this ensemble not only achieved unparalleled accuracy but also emphasised the paramount importance of high recall in medical diagnostics. The delicate balance between specificity and sensitivity, as observed in the ensemble model, offers a profound reflection on the intricate considerations that underpin AI applications in healthcare.

Future Work

The results obtained from our ensemble methods, particularly the Weighted Ensemble 2, demonstrate promising sensitivity, specificity, and accuracy in skin cancer detection. However, when juxtaposed with the findings from other recent studies, there's a discernible room for improvement, especially in the area of Area Under the Curve (AUC) and specificity.

Future research endeavours could focus on:

- Enhanced Preprocessing: While our preprocessing methods, including oversampling, augmentation, and resizing, have been effective, exploring advanced preprocessing techniques might further refine the dataset and enhance model performance.
- Model Fine-tuning: The weights assigned in the ensemble methods can be further optimised using techniques like genetic algorithms or grid search to find the optimal combination.

- Incorporation of Additional Models: Introducing other high-performing deep learning architectures into the ensemble might provide a more diverse set of predictions, potentially enhancing the overall performance.

- Feature Engineering: Extracting and selecting more relevant features from the images might improve the model's ability to discern between benign and malignant lesions.

- Comparison with State-of-the-Art: A deeper analysis of the methodologies employed in the studies that achieved higher AUC values, such as those by Gehad et al. (2021) and Guergueb et al (2021), could provide insights into potential improvements.

In conclusion, while our ensemble approach has shown commendable results, the ever-evolving field of deep learning and dermatological diagnostics offers numerous avenues for further exploration and enhancement.

Appendix

Code link
https://cseegit.essex.ac.uk/ms22045/22-23_ce901-ce901-su_saad_muhammad/-/blob/main/Final_Skin_Cancer_Detection__1_.ipynb

Acknowledgements

I sincerely like to thank Professor Alba Garcia for her valuable suggestions and guidelines to help me complete my dissertation. Moreover, special thanks to Dr. Gehad Islmail, publisher of the paper “A novel melanoma prediction model for imbalanced data using optimised SqueezeNet by bald eagle search optimization”, for providing necessary help in preprocessing skin cancer images.

Bibliography

Rotemberg, V., Kurtansky, N., Betz-Stablein, B., Caffery, L., Chousakos, E., Codella, N., Combalia, M., Dusza, S., Guitera, P., Gutman, D., Halpern, A., Helba, B., Kittler, H., Kose, K., Langer, S., Lioprys, K., Malvehy, J., Musthaq, S., Nanda, J., Reiter, O., Shih, G., Stratigos, A., Tschandl, P., Weber, J. & Soyer, P., 2021. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. Sci Data, 8(1), p.34. Available at: https://doi.org/10.1038/s41597-021-00815-z

World Health Organisation. (Year of publication or last update). Radiation: Ultraviolet (UV) radiation and skin cancer. World Health Organization. Available at: https://www.who.int/news-room/questions-and-answers/item/radiation-ultraviolet-(uv)-radiation-and-skin-cancer. Accessed: 15 August 2023.

Cancer Research UK. (2022). Melanoma skin cancer mortality statistics. Available at: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/melanoma-skin-cancer/mortality#heading-Two (Accessed: 15 August 2023).

Zhang, Y. & Wang, C., 2021. SIIM-ISIC Melanoma Classification With DenseNet. In: 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), Nanchang, China, pp. 14-17.

Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K.Q., 2017. Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708).

Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Zarif, M. and Bowles, R.K., 2020. Mapping diffusivity of narrow channels into one dimension. Physical Review E, 101(1), p.012908.

Sayed, Gehad .I., Soliman, M.M. & Hassanien, A.E., 2021. A novel melanoma prediction model for imbalanced data using optimised SqueezeNet by bald eagle search optimization. Computers in Biology and Medicine, 136, p.104712.

Ghazikhani, A., Yazdi, H. & Monsefi, R., 2012. Class imbalance handling using wrapper based random oversampling. In: 20th Iranian Conference on Electrical Engineering (ICEE2012), IEEE, pp. 611–616.

Barua, S., Islam, M. & Murase, K., 2011. A novel synthetic minority oversampling technique for imbalanced data set learning. In: Neural Information Processing, Springer Berlin Heidelberg, pp. 735–744.

Saini, M. & Susan, S., 2020. Deep transfer with minority data augmentation for imbalanced breast cancer dataset. Appl. Soft Comput., 106759.

Abuared, N., Panthakkan, A., Al-Saad, M., Amin, S.A. & Mansoor, W., 2020. Skin Cancer Classification Model Based on VGG 19 and Transfer Learning. In: 2020 3rd International Conference on Signal Processing and Information Security (ICSPIS), DUBAI, United Arab Emirates, pp. 1-4.

Guergueb, T. & Akhloufi, M.A., 2021. Melanoma Skin Cancer Detection Using Recent Deep Learning Models. In: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Mexico, pp. 3074-3077.

Godlin Jasil, S.P. & Ulagamuthalvi, V., 2021. Skin Lesion Classification Using Pre-Trained DenseNet201 Deep Neural Network. In: 2021 3rd International Conference on Signal Processing and Communication (ICPSC), Coimbatore, India, pp. 393-396.

Khamparia, A., Singh, P.K., Rani, P., Samanta, D., Khanna, A. & Bhushan, B., 2021. An internet of health things‐driven deep learning framework for detection and classification of skin cancer using transfer learning. Transactions on Emerging Telecommunications Technologies, 32(7), p.e3963.

Brinker, T.J. et al., 2019. A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task. Eur J Cancer, 111, pp.148-154.

Guergueb, T. & Akhloufi, M.A., 2022. Multi-Scale Deep Ensemble Learning for Melanoma Skin Cancer Detection. In: 2022 IEEE 23rd International Conference on Information Reuse and Integration for Data Science (IRI), San Diego, CA, USA, pp. 256-261.

Ningrum, D. et al., 2021. Deep learning classifier with patient's metadata of dermoscopic images in malignant melanoma detection. J Multidiscip Healthc, 14, pp. 877-885.

Ms, D., Pyingkodi, M., Thenmozhi, D.K. & Hemalatha, Y.D., 2020. Skin cancer classification towards melanoma detection with deep learning techniques. International Journal of Advanced Science and Technology, 29(9s), pp. 3911-3918.

Sayed, G.I., Soliman, M.M. & Hassanien, A.E., 2021. A novel melanoma prediction model for imbalanced data using optimised squeezenet by bald eagle search optimization. Computers in Biology and Medicine, 136, pp. 104712.