

Can AI Spot Skin Cancer Lesions

A REPORT SUBMITTED TO MANCHESTER METROPOLITAN
UNIVERSITY FOR THE DEGREE OF BACHELOR OF SCIENCE
IN THE FACULTY OF SCIENCE AND ENGINEERING



2023

By

Muhammad Ali Syed

Department of Computing and Mathematics

CONTENTS

Contents.....	2
List of Figures	5
List of Tables.....	7
Abstract.....	8
Declaration.....	9
Acknowledgements	1
1 Introduction.....	2
1.1 Skin Cancer	2
1.1.1 DR ABCDE	3
1.1.2 Ugly Duckling.....	4
1.1.3 Early Detection	5
1.2 Project Aims and Objectives	5
1.2.1 Aim	5
1.2.2 Objectives	5
1.3 Deep Learning	6
1.3.1 History of Deep Learning	6
1.3.2 CNN Layers	7
1.3.3 Activation Functions.....	9
1.3.4 Datasets	10
2 Previous Works	11
2.1 Normal Image Classification	11
2.2 Ugly Duckling	13
2.3 DR ABCDE	15
3 Method.....	16
3.1 Tools.....	16
3.2 Datasets.....	16
3.3 Dataset Preparation	17
3.4 Models	17

3.5	Optimisers	18
3.5.1	STOCHASTIC GRADIENT DESCENT	18
3.5.2	ADAM	18
3.6	Metrics	19
3.6.1	Accuracy	19
3.6.2	AUC	19
3.6.3	Confusion Matrix.....	20
4	<i>Models and results.....</i>	22
4.1	First Batch – Accuracy.....	22
4.1.1	Regularisation	23
4.1.2	Transfer Learning Model.....	25
4.2	Second Batch – AUC	26
4.3	Third Batch – Larger Input Size.....	28
4.4	Fourth Batch – Categorical Labels	29
4.4.1	One Hot Encoding	29
4.4.2	Models and Results for One Hot encoded	31
4.5	Fifth batch – Smaller Batch Size	32
4.6	Conclusion	33
5	<i>Ugly Duckling.....</i>	34
5.1	Autoencoder	34
5.2	Preparation.....	36
5.3	Dataset	36
5.4	Training	37
5.5	Anomaly Detection	38
5.6	Conclusion	39
6	<i>Evaluation</i>	41
6.1	Single Image Classification	41
6.2	Ugly Duckling And Dr ABCDE.....	42
7	<i>Conclusion And Reflection</i>	43
7.1	Future Work.....	43
8	<i>References</i>	45

9	<i>Appendix – Show Case Slides.....</i>	49
----------	--	-----------

LIST OF FIGURES

Figure 1. Five main stages of Melanoma (Source: People Beating Cancer, 2022:online)	3
Figure 2. ABCDE images (Source: Skin Cancer Foundation, 2021:online)	4
Figure 3. Three examples of Ugly Ducklings (Source: Skin Cancer Foundation, 2021:online)	5
Figure 4.Example of a filter applied to a Two-Dimensional Input to create a Feature Map. (Source: Brownlee, 2019:online)	7
Figure 5. Examples of Max Pooling and Average Pooling.....	8
Figure 6. Example of a fully connected layer (Source: Sharma, 2020:online).....	8
Figure 7. Graph of a Sigmoid Function	9
Figure 8.Visualisation of the ReLU function	10
Figure 9.Ugly Duckling Detection Pipeline (Source: Mohseni et al., 2021)	14
Figure 10. Visualising performance of AUC curves. (Source: Wikipedia).	20
Figure 11. Baseline Model architecture.	23
Figure 12. Architecture of Baseline model with dropout layers.	24
Figure 13. Graph of Cross Entropy Loss against epochs for the Baseline Model with Image Augmentation with orange for validation and blue for training loss.	24
Figure 14. Structure of the VGG16 Model.....	25
Figure 15. Probability Distribution of predicted labels returned by the model.....	30
Figure 16. Confusion matrix for predictions before one hot encoded labels	30
Figure 17. Probabilities returned by mode, first column for Benign and second for Malignant.	30
Figure 18. Structure of an Autoencoder (Source: Dertat, 2017:online).....	35
Figure 19. Individual Gaussian kernels plotted along with the Kernel Density Estimate. (Source: Lets talk About Science, 2020:online).....	36
Figure 20. Lesion image before and after segmentation	37
Figure 21. Input images and corresponding outputs	37
Figure 22. Confusion matrix for the predictions.....	38
Figure 23. Reconstruction errors and standard deviations for Benign and Malignant Images.....	39

Figure 24. MSE loss graph for the model.....	39
--	----

LIST OF TABLES

Table 1. Showing the Breakdown of the 3 datasets used.	17
Table 2. Confusion matrix for the ISIC dataset	21
Table 3. Shows results for the different models with the highest values of AUC performance metric during training and loss graphs.	27
Table 4. Results for 9 models with 224x224 input size and Adam optimiser.....	28
Table 5. AUC results and confusion matrices from fourth batch of models.	31
Table 6. AUC and Confusion matrices for fifth batch of models.	33
Table 7. Breakdown of the datasets to be used for Training and Testing of the Autoencoder.	37

ABSTRACT

This report explores the use of Neural Networks in detecting skin cancer lesions.

Skin Cancer at its early stages is highly curable, however the survival rate drops significantly if the condition is diagnosed late. Dermatologists use methods called the Ugly Duckling method, where a skin lesion is compared to other skin lesions present on a patient's body and anomalous lesions are flagged for excision to diagnose for skin cancer. Another method called the DR ABCDE, where a dermatologist looks at the appearance of a lesion to diagnose.

This research focusses at investigating current use of Neural Networks to diagnose skin cancer and potential implementation of the rarely used Ugly Duckling and DR ABCDE method for Neural Networks.

DECLARATION

No part of this project has been submitted in support of an application for any other degree or qualification at this or any other institute of learning. Apart from those parts of the project containing citations to the work of others, this project is my own unaided work. This work has been carried out in accordance with the Manchester Metropolitan University research ethics procedures and has received ethical approval number 52607.

Signed: *Muhammad Ali Syed*

Date: 19th May 2023

ACKNOWLEDGEMENTS

I would like to express my heartfelt appreciation and gratitude to my supervisor, Connah Kendrick, for their invaluable guidance, unwavering support, and insightful feedback throughout the duration of this project. Their expertise and willingness to share their knowledge have been instrumental in helping me understand the intricacies of the project, enabling me to navigate through challenges effectively.

I am truly grateful for the patience and encouragement provided by my supervisor, as they consistently dedicated their time and effort to address any queries or concerns I had. Their guidance not only clarified complex concepts but also fostered my personal and professional growth. Their mentorship played a pivotal role in shaping the successful outcome of this report.

Additionally, I would like to extend my sincere appreciation to Nicholas Costen, for their invaluable contributions to this project. Their assistance in identifying and resolving critical issues in my product was truly invaluable. Their profound knowledge and dedication to ensuring excellence inspired me to strive for continuous improvement. Furthermore, their invaluable insights and suggestions played a significant role in structuring this report effectively.

1 INTRODUCTION

1.1 SKIN CANCER

The skin, as the body's largest organ, plays a vital role in protecting the internal structures from damage. Its primary defence against harmful UV radiation from the sun, which can cause damage to the genetic material (DNA) of cells and appear as a sunburn on the skin (Cancer Research UK, 2020), is facilitated by melanocytes, a group of cells residing deep within the epidermis. These melanocytes produce melanin, a pigment that absorbs UV radiation and reduces its penetration into the skin layers (Brenner and Hearing, 2008). While melanocytes have a built-in kill switch mechanism designed to eliminate damaged cells, occasionally, this fails to function properly, allowing the survival and multiplication of damaged cells. This aberration can lead to the development of melanoma, a type of skin cancer (ScienceDaily, 2010).

As depicted in Figure 1 (People Beating Cancer, 2022:online), Melanoma inside the skin develops in five main stages. At stage 0, the cancer cells reside in only the outermost layer of the skin, the epidermis. Diagnosis at this stage has almost a 100% survival rate for 5 years or more. Stage 1 and Stage 2 occur when the cancer cells first start to spread into the dermis layer as well and over time get bigger in size. Diagnosis at these stages has a lower survival rate of 70% - 80%. At Stages 3 and 4, the cancer cells spread to the regional lymph nodes and using these pathways spread to other parts of the body including the brain, at which point the cancer becomes extremely challenging to treat and survival rates of individuals significantly decrease to around 30%. (Aim At Melanoma, 2023)(Cancer Research UK, 2020).

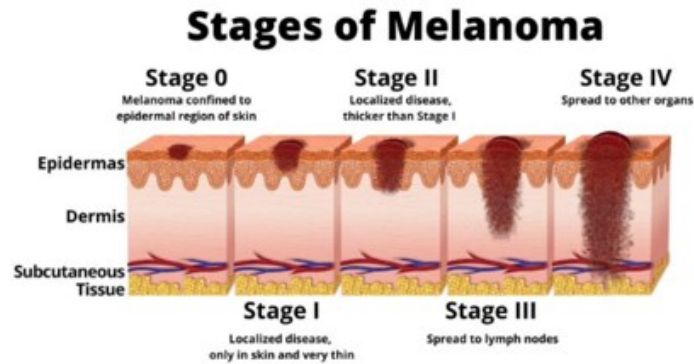


Figure 1. Five main stages of Melanoma (Source: People Beating Cancer, 2022:online)

Around 16,700 people are diagnosed with Melanoma each year in the UK (Cancer Research UK, 2016-2018) and almost all of these cases are easily curable by simply removing the cancerous cells if the cancer is detected and diagnosed at its earlier stages, which signifies the importance of early detection. Currently Dermatologists employ a host of manual techniques which can help in determining if a skin lesion could be melanoma or nevi.

1.1.1 DR ABCDE

These techniques include the DR ABCDE method where the Dermatologists make use of a specialised handheld instrument called a dermatoscope, which enables the skin specialist to see the finer details of a skin lesion (Cancer Research UK, 2020). Looking at these details of the lesion, the dermatologist checks for key “ABCDE” features of the skin lesion as illustrated in Figure 2 (Skin Cancer Foundation, 2021:online).

- A for Asymmetry: One half of the skin lesion may not look like the other half.
- B for Border: The border of the skin lesion may be irregular.
- C for Colour: The skin lesion may have varying colours.
- D for Diameter: The lesion may be larger than 6mm.
- E for Evolving: The spot changes size, colour, shape, etc. over time.

(American Academy of Dermatology Association, 2023).

Finding abnormalities in any of these features gives a sign that the skin lesion may be malignant, and a skin biopsy is performed where the skin lesion is removed and examined under a microscope in a lab. If the invasive procedure proves the presence of malignant cells, deeper

tissue removal is performed which is often enough to cure the cancer without further treatment if done in the early stages (American Cancer Society, 2019).

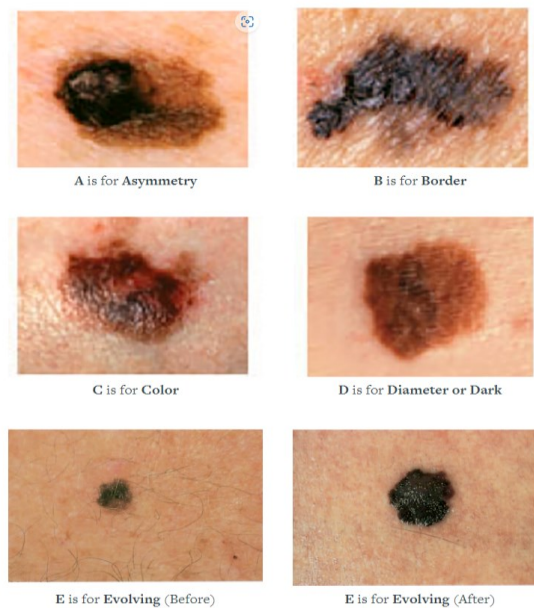


Figure 2. ABCDE images (Source: Skin Cancer Foundation, 2021:online)

1.1.2 Ugly Duckling

Another more commonly used technique is the Ugly Duckling method where the dermatologist takes all the skin lesions present on a patient's body into account and based on the concept that most nevus on an individual's body resemble one another, while a melanoma stands out and looks different in comparison, make a diagnosis based on their judgment and experience. As illustrated in Figure 3 (Skin Cancer Foundation, 2021:online), an ugly duckling may be larger, smaller, lighter, or darker compared to neighbouring lesions making them stand out and signal the presence of melanoma, similarly to the ABCDE method, this is also very subjective, and diagnosis is made depending on the medical professional's knowledge and experience.

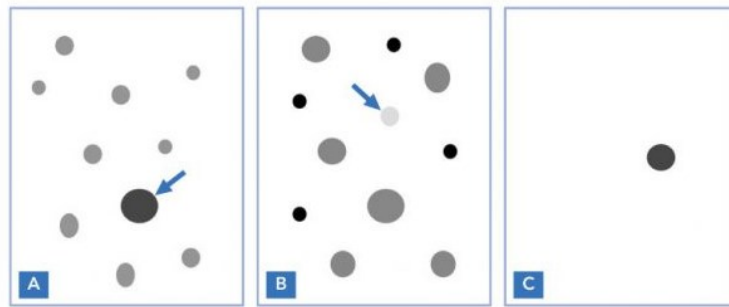


Figure 3. Three examples of Ugly Ducklings (Source: Skin Cancer Foundation, 2021:online)

1.1.3 Early Detection

The immense importance of early detection in skin cancer diagnosis renders a false negative diagnosis detrimental to the survival rates of the patients, leading to medical professionals diagnosing a very large number of false positives in the name of caution. This results in unnecessary excision surgeries that put major avoidable stress on medical facilities. Diagnostic accuracies currently are as low as 8% (Grove, Green, 2020) and the reason is that the medical professionals that carry out most of the diagnoses are general practitioners and lack the in-depth Dermoscopy training which is either expensive and time-consuming or inaccessible in poor areas or more generally, a professional dermatologist is required to make a diagnosis, which people don't have access to in many cases. This further strengthens a need for an alternative, more accessible and higher accuracy diagnosis tool, which would decrease the need for unnecessary skin biopsy procedures, reducing the unnecessary stress on medical facilities and resources.

1.2 PROJECT AIMS AND OBJECTIVES

1.2.1 Aim

The aim of this project is to analyse the use of Neural Networks in skin cancer detection and research the possible implementation of the Ugly Duckling and DR ABCDE methods for Neural Networks.

1.2.2 Objectives

- Research and test current skin cancer detection methods
- Investigate the Ugly Duckling method

- Investigate the 7-point checklist for the DR ABCDE method
- Analyse the performance and evaluate findings in comparison to previous research

1.3 DEEP LEARNING

A major step in the efficient diagnosis of skin cancer lesions has been the use of Deep Learning. Deep learning is a subset of machine learning where computer algorithms are designed to learn and emulate human-like intelligence. By processing vast amounts of data, they perform tasks such as image classification, language translation, speech recognition etc. Deep learning neural networks are layers of nodes similar in function to neurons in a human brain where these nodes receive signals from previous layers in the network, apply weights to this signal and pass them down to deeper layers. Higher weights result in a greater effect on the following layers and the final layers produce output depending on this weighted input. Because of the large amount of data and complex mathematical calculations, deep learning neural networks require powerful hardware which has only recently been made possible due to advancements in Dig Data analytics which overcame these computing power limitations and allowed multi-layered neural networks to be trained on very large amounts of data, giving rise to deep learning. (IBM, 2023)

1.3.1 History of Deep Learning

The history of deep learning can be traced back to 1943 when Walter Pitts and Warren McCulloch created a computer model which used a combination of algorithms and mathematics called “threshold logic” based on the neural networks of the human brain (Foote, 2022). Around the 1960s, Henry J. Kelly developed this into a basic Back Propagation model (Foote, 2022) which forms the essence of neural network training where a neural network during training, takes the error rate of forward propagation and feeds this loss backwards through the neural network layers to fine-tune the weights, aiming to reduce this loss (Al-Masri, A. 2022).

Further research lead Kunihiko Fukushima to develop the first convolutional neural network with multiple pooling and convolutional layers in the 1970s (Foote, 2022). This design, allowing computers to “learn” to recognize visual patterns, formed a basis for the further development of convolutional neural networks over the next few decades. LeNet, created in 1998 by Yann LeCun was the first pre-trained CNN architecture to employ back-propagation in a practical

application, giving birth to deep learning. The architecture, although very simplistic, consisting of just 5 layers, outperformed all other models at the time in handwritten digit recognition challenges (Ganesh, P. 2020) and paved the way for more complex models.

Convolutional Neural Networks (CNNs) play an important role in deep learning due to their ability to efficiently process and analyse large amounts of data with a grid-like structure i.e. an image. Mainly consisting of convolutional, pooling and activation layers, CNNs are capable of detecting complex patterns in data and making predictions or classifications. We will talk more about these layers and their importance in CNNs.

1.3.2 CNN Layers

The convolutional layer, as the name suggests, performs an operation called a “convolution” on an input image to extract useful features and patterns. This involves sliding a matrix, called a kernel or filter, over the image to perform multiplication and summation, resulting in an output, called a feature map as illustrated in Figure 4 (Brownlee, 2019). This feature map represents the presence or absence of specific features in an image, regardless of the location of the feature in the image. This matrix is also called the weights for a convolutional layer, which a CNN is capable of learning on its while training by back-propagating and adjusting these matrix values. (Brownlee, 2019)

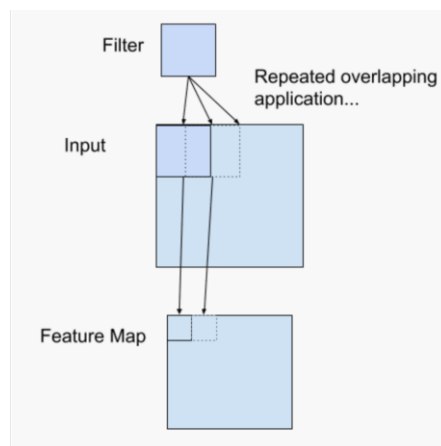


Figure 4. Example of a filter applied to a Two-Dimensional Input to create a Feature Map. (Source: Brownlee, 2019:online)

The Pooling layer, present after the convolutional layer also performs dimensionality reduction by extracting dominant features from the output of the previous layer. There are two types of

pooling layers, the Max Pooling layer where the maximum value is returned from the portion of the image covered by the kernel and the Average Pooling layer where the average of all the values covered by the kernel is returned as illustrated in Figure 5. Both layers perform dimensionality reduction but Max Pooling acts as a de-noising mechanism as well which makes it better than its counterpart. In deep learning, due to the large amount of computation necessary, this dimensionality reduction is required to build efficient models. (Saha, 2018).

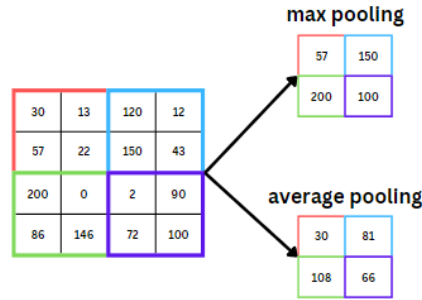


Figure 5. Examples of Max Pooling and Average Pooling

Finally, the Dense Layer, also known as the fully connected layer because every neuron in this layer receives inputs from all neurons in the previous layer as illustrated in Figure 6 (Sharma, 2020:online), produces an output for each image depending on some learned parameters from the training process. Each neuron in the dense layer applies a non-linear activation function to the inputs, allowing the model to learn complex patterns in the images. The two most used activation functions are the Rectified Linear Unit (ReLU) function and the sigmoid function (Verma, 2021).

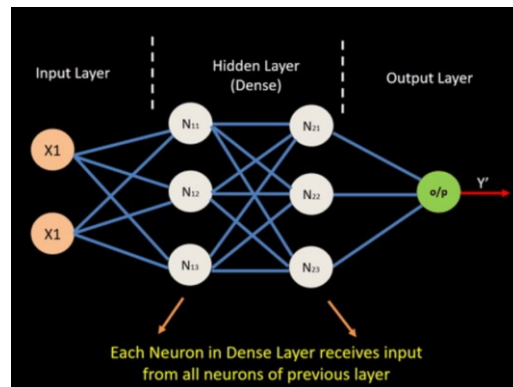


Figure 6. Example of a fully connected layer (Source: Sharma, 2020:online)

1.3.3 Activation Functions

As shown in Figure 7, The sigmoid function is an S-shaped function that outputs in the range of zero and one, setting all inputs below zero to zero and above one to one. Although common, this activation function is only sensitive around its mid-point and causes saturation due to snapping very large and very small values to zero and one. Once saturated, it becomes very difficult for deep learning algorithms to continue to adapt the weights to improve the performance resulting from the vanishing gradient problem. The vanishing gradient problem is where when backpropagating towards the earlier layers to update the weights to reduce the loss function, gradients are computed with respect to the weights of each layer. This gradient can get exponentially small to the point that rate of convergence slows down resulting in a lack of significant contribution of earlier layers towards learning patterns resulting in decreased performance (Brownlee, 2019).

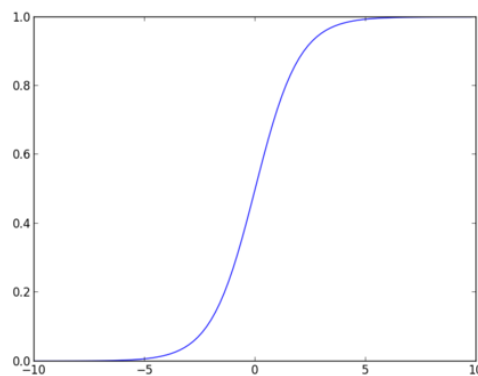


Figure 7. Graph of a Sigmoid Function

To fix this problem, a function was required that would mimic a linear function but still allow for complex relations to be learned from the data. This was the ReLU function (Figure 8), which is a piecewise linear function that sets all values below zero to zero but keeps values above zero as they are, resulting in more sensitivity and preventing saturation. Due to its significance in training deep learning models efficiently, ReLU is regarded as one of the few milestones in the deep learning revolution. (Brownlee, 2019)

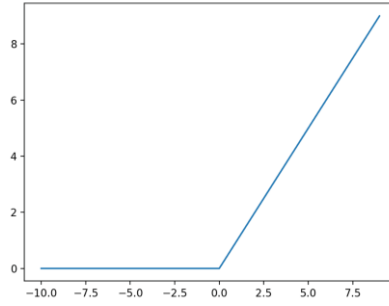


Figure 8. Visualisation of the ReLU function

1.3.4 Datasets

The results of a neural network are only as good as the dataset that is used to train them, so immense importance is placed upon the availability of high-quality and professionally annotated datasets. The International Skin Imaging Collaboration (ISIC) is an academic and industry partnership that was created keeping this need in mind and is currently creating resources for the dermatology and computer science communities, including a large and expanding open-source public access archive of around 70,000 standardised high-quality skin images, that serve in the development and testing of diagnostic artificial intelligence algorithms (ISIC, 2023).

Another large dataset of around 17,000 expertly annotated images is the Fitzpatrick 17k which was sourced from two dermatology atlases, the DermaAmin and the Atlas Dermatologico. This dataset was annotated based on the Fitzpatrick scoring system to evaluate algorithmic fairness and model accuracy in computer vision (Groh, 2022). Additionally, there are also a few smaller datasets available such as Derm7pt with over 2000 images, Dermofit Image Library with around 1,300 images and PAD-UFES-20 with over 2,200 images (ISIC, 2022). Although these datasets are professionally annotated and meet all the criteria for being high-quality datasets for NN training, they are far too small to be used for large deep-learning models and usually require more images from other sources to be added.

2 PREVIOUS WORKS

In this section, we will be discussing the current research revolving around the use of Convolutional Neural Networks in image classification to classify skin lesions. All research papers we have come across reiterate the importance of developing Neural Networks that can diagnose skin cancer lesions more accurately than dermatologists that are often hard to train or inaccessible to most people.

2.1 NORMAL IMAGE CLASSIFICATION

In “Skin Cancer Classification Using Convolutional Neural Networks: Systematic Review”, Brinker et al. (2018) review 13 research papers that used Convolutional Neural Networks (CNN) to classify skin lesions. Their investigation of the results from these papers showed that CNNs are very capable of achieving high accuracy, reportedly even as high as 98.7%, in classifying skin lesions. However, Brinker et al. also identified major challenges that would have to be addressed before CNNs can be implemented in clinical settings. These included the need for larger and more diverse datasets, which in their reviewed papers, were relatively very small and would not be able to generalise well on new data. Another challenge was that the model structures were relatively simple and more complex model architecture could see potential improvement in performance. Finally, clinicians want to understand how CNNs reach their decisions to trust their results but currently, due to a lack of interpretability in CNNs, this raises a problem. Addressing these problems in the future could see a rise in the clinical use of CNNs.

In “Analysis of the ISIC image datasets: Usage, benchmarks and recommendations” Cassidy et al. (2022) analysed the usage of the ISIC datasets from a period of 2016 -2020 and discussed challenges concerning these datasets which include the presence of duplicate images, image similarity and the imbalance of classes in the datasets, highlighted the importance of the reliability and consistency of open-source datasets currently present for the development of algorithms. According to their survey of several well-cited reports, almost all the reports did not implement any duplicate removal strategy which undermined the integrity of the results. They proposed an image duplicate removal strategy to “clean” datasets and allow for a more accurate measure of model performance. Benchmarking this new “cleaned” dataset, containing no

duplicate images and an equal number of images for each class, on 19 deep learning architectures and evaluated on the testing dataset from ISIC 2020 and ISIC 2017.

In “Melanoma and Nevus Classification using Convolution Neural Networks”, Grove and Green (2020) propose a method of identifying malignant skin lesions from images taken at the general practice level. Building on previous research that used image pre-processing in the form of filtering, edge detection and colour distribution analysis, Grove and Green propose a method that removes these complex series of geometric algorithms and replaces them with a single convolutional neural network. The Resnet50 CNN, which has 50 layers and is trained on more than a million images from the ImageNet database (MathWorks, 2023) was made use of a transfer learning approach to reduce the training efforts of the ISIC dataset.

The overall results were good with an accuracy of 86% but the model particularly struggled with some weirdly shaped nevus that were misclassified as melanomas. Furthermore, the network was unreliable when classifying nevus that did not possess the dramatic colours of the most easily identifiable melanoma lesions. Grove and Green also proposed some ideas to address the limitation of the network, by altering the training dataset to allow the network to classify malignant lesions in addition to melanomas and implementing a different CNN to compare accuracies. Another major idea for improvement presented is to extend the network to allow the classification of not only dermoscopic images, limiting the use to medical professionals but also images from portable devices, making them more accessible.

Bi et al. (2017) in the paper “Automatic Skin Lesion Analysis using Large-scale Dermoscopy Images and Deep Residual Networks” and Acosta et al. (2021) in their paper “Melanoma Diagnosis using deep learning techniques on dermatoscopic images”, proposed two different methods of automatic skin lesion analysis on the ISIC 2017 dataset with 2000 images. Acquiring 8000 additional images from the ISIC archive, Bi et al. used a transfer learning approach by finetuning a ResNet50 pre-trained model to extract features from the images and then using three classification approaches, multi-class classification, binary classification, and an ensemble approach, evaluate the performance of the method. Their proposed method achieved first place for the validation set in the ISIC 2017 challenge with a Jaccard of 79.40%.

On the other hand, Acosta et al. use a different approach where they first use a feature-extracting model called a Mask and Region-based Convolutional Neural Network (Mask R_CNN) to detect a skin lesion in the image and crop a bounding box around the skin lesion. This cropping is done to overcome the difficulties in classification caused by the presence of hairs, inks, ruler markings, coloured patches etc. in images. Then, using a finetuned ResNet152, make a classification on this cropped lesion image. Their method managed to achieve a high specificity and sensitivity score of greater than 0.8.

2.2 UGLY DUCKLING

Outlier detection, anomaly detection or one-class classification which forms the basis of the Ugly Duckling method, remains a major unsolved challenge in deep learning. Many researchers have attempted to present their solutions for the challenge, some of which we will be discussing below.

As discussed previously, deep learning models require an immense amount of data to train to an acceptable level of accuracy. This large amount of data required must be expertly annotated which is often hard to access, very expensive or time-consuming. In “Variational Autoencoder based Anomaly Detection using Reconstruction Probability”, An and Cho (2015) present an attempt to solve the outlier detection problem, which does not require labelled datasets and can be trained with unlabelled datasets that are readily available.

An and Cho (2015) propose using a Variational Autoencoder to return reconstruction probability for images and using these to determine anomalies. Reconstruction probability is different from reconstruction error, which is returned by normal Autoencoders, in the sense that reconstruction probability is the measure of how likely it is that a data point was generated by a VAE, considering the distribution of the data. Conversely, reconstruction error is just a measure of how close the output is to the input data, considering only the distance between the data points. This difference meant that VAEs are more sensitive than AEs and were chosen by An and Cho (2015) to produce their method of anomaly detection. Their method involved training a VAE on a normal dataset of normal points and then using this VAE to produce reconstruction probabilities for test images. A data point having a low reconstruction probability would be classified as an

anomaly. Evaluating the MNIST digits datasets, they were reportedly able to achieve an accuracy of 99%.

In “Can self-training identify suspicious ugly duckling lesions?”, Mohseni et al. present a deep learning model that attempts to build on the earlier outlier detection solution presented for the Ugly Duckling method by making use of a deep learning technique called self-training. Self-training is a semi-supervised machine learning technique which allows the use of unlabelled dataset which is more readily available. A larger dataset usually correlates to higher model accuracy.

In Mohseni et al.’s method, a multi-staged pipeline was developed as illustrated in Figure 9 (Mohseni et al., 2021) where the first module detects, and extracts skin lesions present on a TBP (Total Body Photography) image. The next module segments the skin lesions to be used in the outlier detection module. The outlier detection module makes use of a neural network called a Variational Autoencoder (VAE) which learns a low-dimensional representation of the input data. VAEs try to reconstruct common-looking lesions which are the majority, as perfectly as possible but Ugly Ducklings which are a minority are less prioritised leading to a higher reconstruction loss so defining a threshold on the reconstructed losses, lesions were classified as common or UD.

The model produced promising results with an average accuracy of 94%. But the model accuracy remains limited to the availability of high-resolution images that may be hard to acquire from a handheld camera device. Further work is required to make images from mobile devices where details of very small lesions, less than 1.5mm, may be lost due to the lower resolution.

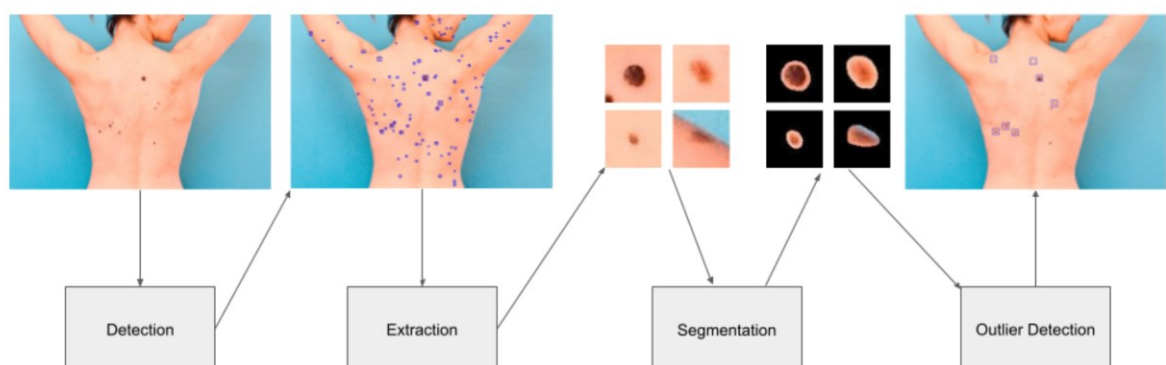


Figure 9. Ugly Duckling Detection Pipeline (Source: Mohseni et al., 2021)

2.3 DR ABCDE

Not much research is available for the 7-point checklist implementation using neural networks for the DR ABCDE method and Kawahara et al. (2019) highlight this limitation in their paper titled “Seven Point Checklist and Skin Lesion Classification Using Multitask Multimodal Neural Nets”, where they discuss the existing work that solely focuses on implementing only one criterion of the checklist. Their proposed method involved making predictions using a multimodal Convolutional Neural Network that would consider multimodal data in the form of clinical images, dermoscopic images and patient metadata while training, using multiple loss functions, handling the combinations of these input modalities. This model then produces a multimodal feature map, capable of classifying images according to the 7-point checklist.

3 METHOD

3.1 TOOLS

TensorFlow, which is an open-source library for large-scale machine learning, created by the Google Brain team in 2015 was used with Python language to train and run deep neural networks for image classification (Yegulalp, 2022). The main reason for using TensorFlow was because it provides an abstraction to the machine learning development process. Instead of dealing with the details of implementing an algorithm for example figuring out the proper way of inputting an output of one function to another function, TensorFlow allowed the focus to be on the overall application logic and handled all the details behind the scenes. Google Collaboratory, which is a cloud-based IDE (Integrated Development Environment) providing free access to CPU and GPU hardware, was used to write and execute the Python code through the browser. This access to high-spec hardware made it particularly important in the development of this project.

3.2 DATASETS

Three sub-datasets of the ISIC 2017 dataset were used in the development of the project. The largest of the three, the training dataset, allowed the models to learn features and patterns in the images during the training phase. Next, the validation dataset, which forms a crucial part of the training phase, is an unseen dataset that after each epoch, evaluated the performance of the models. This performance was compared to the training dataset's performance and any differences in these performances suggested that for example when the validation dataset was performing worse than the training dataset, the model was finding it difficult to generalise over new and unseen data so it may be overfitting and conversely, when both datasets were performing poorly, the model was struggling to capture complex patterns in the data and may be underfitting. These observations suggested the need for adjustments to be made to the models to improve performance. Finally, another unseen dataset, called the test dataset, was used to provide an unbiased evaluation of the performance of the deep learning models to ensure models can perform effectively in a real-world situation.

3.3 DATASET PREPARATION

We used training, validation and testing datasets from the ISIC 2017 website which contained 2000, 150 and 600 dermoscopic skin lesion images respectively as shown in Table 1. These datasets were chosen for the project as they were expertly annotated and large enough to train deep learning models efficiently but not too large to the point that it hinders performance and training times due to a lack of access to high-spec hardware required for quicker training times.

Dataset	Benign Images	Malignant Images	Total
Train	1626	374	2000
Validation	120	30	150
Test	483	117	600

Table 1. Showing the Breakdown of the 3 datasets used.

The datasets were downloaded from the official ISIC challenges website with the corresponding training, validation and testing ground truth CSV (Comma Separated Values) files which contained diagnosis for each image. With the help of the ground truth CSV file containing labels for each image, the images in the corresponding datasets were further divided into two new directories, one containing only Benign skin lesion images and the other containing only Malignant skin lesion images. The datasets also contained superpixel images, but these were deemed insignificant and removed from the final datasets for simplicity.

3.4 MODELS

During the development of this project, only sequential models were used to train on the datasets. Sequential models are a type of neural network that have layers that are stacked sequentially, one after the other, forming a linear pipeline where the input data gets fed at one end and after applying non-linear transformations in the hidden layers to this data, an output is produced from the final layer. These models included custom CNNs as well as pre-trained CNNs such as VGG16, VGG19 etc.

All the images before being passed to the models were standardised by re-scaling the pixel values to be between 0-1 with the help of ImageDataGenerator. ImageDataGenerator is a utility in Keras that gives the ability to apply augmentations to images on-the-fly. It also contains a method called “flow from directory” which was used to read the images from the specific directories, apply pre-processing and prepare the images to be passed to the models’ fit, evaluate and predict methods. An initial batch size of 64 and binary class mode were specified for the initial batches of models, but these were later changed after the evaluation of the result, discussed further in the next section.

3.5 OPTIMISERS

An optimiser is an algorithm that during training of a model, takes the gradients of the loss function and updates the parameters of a model, trying to decrease this loss to find the best parameters. Mainly, two optimizers were used in the development of the project. Initial models employed the SGD (Stochastic Gradient Descent) optimiser, but later models switched to the Adam optimiser, reasons are discussed in the following section.

3.5.1 STOCHASTIC GRADIENT DESCENT

SGD is an optimiser which unlike traditional gradient descent algorithms that take the gradient of the loss function for the entire training dataset with respect to the parameters, takes the gradient of the loss function for a smaller subset, called a mini batch, of the training dataset and takes more frequent steps towards minimising the loss. This is faster and more computational efficient than traditional gradient descent, making it especially useful in deep learning where datasets are very large. Learning rate defines how big the steps SGD would take towards decreasing the loss. (Scikit learn, 2023)

3.5.2 ADAM

Adam, which stands for “Adaptive Moment Estimation”, is another optimiser that is used in the training phase of a model. Adam is an extension of the SGD optimiser but is different in the sense that unlike in SGD where the learning rate stays constant throughout the training, Adam computes separate learning rates for each parameter in the model, based on previous gradient information. This allows Adam to automatically adjust the learning rates for each parameter,

leading to faster convergence and better performance. This ability to have variable learning rates makes Adam more suitable in deep learning than SGD. (Brownlee, 2017)

3.6 METRICS

Metrics are used to evaluate the performance of a model and compare with other models. In the development of this project, mainly two metrics were used, initially Accuracy but later switched to AUC (Area Under the ROC (Receiver Operator Characteristic) Curve).

3.6.1 Accuracy

Accuracy is a performance metric that evaluates the fraction of correct predictions over the total number of predictions shown in a mathematical formula in Figure 1. For a binary classification problem, it is calculated by adding the True Positive and True Negative predictions and dividing by the total number of predictions. However, it can produce deceptively good performance scores for datasets where one class may be underrepresented for example if a dataset contains 80% of one class and the model predicts all the images in that dataset to be of that class, the accuracy metric gives us a score of 80% which in reality is not a good score because of the lack of correct classifications of the underrepresented class. Initial models of the project employed accuracy as the performance metric but were soon switched to a better alternative as this limitation was experienced.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

Equation 1. Mathematical formula to calculate Accuracy

3.6.2 AUC

A better alternative to the Accuracy metric for imbalanced datasets is the AUC (Area Under the ROC (Receiver Operator Characteristic) Curve) metric which measures the ability of the model to distinguish between positive and negative classes irrespective of their proportions in the dataset. ROC curve is a plot of the True Positive Rate, which is the ratio of the correctly classified positive sample to the total number of positive samples, against the False Positive Rate, which is the ratio of the incorrectly classified negative samples to the total number of samples (Equation 2.). The area under this curve provides an evaluation of performance as shown in Figure 10, with

a value ranging between zero and one. A higher value indicates that the model was good at differentiating between the two classes and conversely if for example, the value is 0.5, it suggests the model is just randomly guessing between the two classes. (Google Developers Machine Learning, 2022)

True Positive Rate (TPR)

$$TPR = \frac{TP}{TP + FN}$$

False Positive Rate (FPR)

$$FPR = \frac{FP}{FP + TN}$$

Equation 2. Mathematical formula for TPR and FPR.

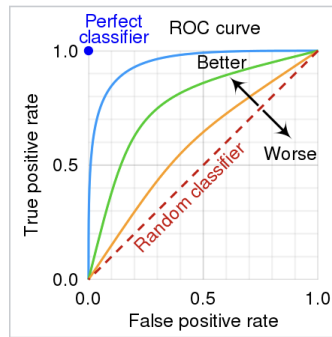


Figure 10. Visualising performance of AUC curves. (Source: Wikipedia).

The following performance equivalences for AUC provide a framework to assess discrimination levels:

- 0.5 = No discrimination
- 0.5 – 0.7 = Poor discrimination
- 0.7 – 0.8 = Acceptable discrimination
- 0.8 – 0.9 = Excellent discrimination
- > 0.9 = Outstanding discrimination

3.6.3 Confusion Matrix

Another evaluation metric adopted for later models in conjunction with the AUC metric was the confusion matrix. A confusion matrix visualises and summarises the performance of a classification algorithm by presenting the results of the model in a tabular form as depicted in Table 2. The columns contain the predicted labels for the images and the rows contain the actual labels. A good-performing model should have most values in a diagonal in the matrix with most predictions being True Positives and True Negatives.

Confusion Matrix	Predicted Benign	Predicted Malignant
True Benign	TN	FP
True Malignant	FN	TP

Table 2. Confusion matrix for the ISIC dataset

4 MODELS AND RESULTS

4.1 FIRST BATCH – ACCURACY

The first batch of models were prepared to give a baseline performance evaluation to compare with future models. These models were also trained and validated using only the 2000 image training dataset in an 8:2 split for simplicity and ease of use at the time. Starting off by following a deep learning tutorial by Brownlee, J. (2019) on the Dogs vs Cats dataset, a baseline model was created with the general architectural principles of the VGG models, as their modular structure is simple enough to understand and was a good starting point. The VGG model architecture involves stacking convolutional layers with small 3×3 filters followed by max pooling layers forming a block. These blocks are stacked on top of one another with increasing number of filters and in each successive block, the number of filters is increased (Brownlee, 2019).

The baseline model consisted of three VGG style blocks with filter sizes of 3×3 in each convolutional layer and the number of filters doubling in each successive layer, starting from 32 filters as illustrated in Figure 11. Each layer in the model used the ReLU activation function and the `he_uniform` initialiser. `he_uniform` initialiser sets all the initial weights in a uniform distribution within a particular range depending on the number of input and output neurons in the layer. This helps prevent the gradient from vanishing or exploding due to weights being initialised to a too high or too low value, making it particularly useful in deep learning neural networks. Each convolutional layer was followed by a max pooling layer with a filter size of 2×2 each. Finally, a sigmoid activation function was used to make the binary classifications between malignant and benign classes. The model was compiled with an SGD (Stochastic Gradient Descent) optimiser with an initial learning rate of 0.001 and 0.9 momentum. An optimiser adjusts the parameters of a model during training to try and decrease the loss function. This decrease in the loss means that the model is performing well, and the optimiser has found the best parameters for the model. Binary Crossentropy was the choice of loss function as the dataset contained only two classes and Accuracy was chosen as the performance metric.

Model: "sequential"		
Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 50, 50, 32)	896
max_pooling2d (MaxPooling2D)	(None, 25, 25, 32)	0
conv2d_1 (Conv2D)	(None, 25, 25, 64)	18496
max_pooling2d_1 (MaxPooling2D)	(None, 12, 12, 64)	0
conv2d_2 (Conv2D)	(None, 12, 12, 128)	73856
max_pooling2d_2 (MaxPooling2D)	(None, 6, 6, 128)	0
flatten (Flatten)	(None, 4608)	0
dense (Dense)	(None, 128)	589952
dense_1 (Dense)	(None, 1)	129
Total params: 683,329		
Trainable params: 683,329		
Non-trainable params: 0		

Figure 11. Baseline Model architecture.

The model was trained for 20 epochs with an image input size of 50*50*3. The reason for choosing a very small input size was to allow for faster training. The model produced promising results with an accuracy of 76.349% on the test dataset which would be used to compare performance of other models.

4.1.1 Regularisation

Regularisation was added to the model to see if a performance gain could be achieved. Building upon the original baseline model, one model was created with added dropout layers as shown in the model architecture in Figure 12. These dropout layers randomly set 20 percent of input units to zero during training to prevent the model from overfitting and allow for it to generalise better on unseen data. Another model was created with image augmentation regularisation added to the input data. Image augmentation artificially grows the size of the training set by creating transformations of the images such as rotating, scaling, flipping etc. This larger training set allows for the model to train on more data, helping recognise more complex patterns.

Model: "sequential_1"		
Layer (type)	Output Shape	Param #
conv2d_3 (Conv2D)	(None, 50, 50, 32)	896
max_pooling2d_3 (MaxPooling 2D)	(None, 25, 25, 32)	0
dropout_1 (Dropout)	(None, 25, 25, 32)	0
conv2d_4 (Conv2D)	(None, 25, 25, 64)	18496
max_pooling2d_4 (MaxPooling 2D)	(None, 12, 12, 64)	0
dropout_2 (Dropout)	(None, 12, 12, 64)	0
conv2d_5 (Conv2D)	(None, 12, 12, 128)	73856
max_pooling2d_5 (MaxPooling 2D)	(None, 6, 6, 128)	0
dropout_3 (Dropout)	(None, 6, 6, 128)	0
flatten_1 (Flatten)	(None, 4608)	0
dense_2 (Dense)	(None, 128)	589952
dropout_4 (Dropout)	(None, 128)	0
dense_3 (Dense)	(None, 1)	129
Total params: 683,329		
Trainable params: 683,329		
Non-trainable params: 0		

Figure 12. Architecture of Baseline model with dropout layers.

It was expected that these regularisation techniques would improve the model's performance but surprisingly both models produced the same results of 76.349% as the baseline model. Further evaluation of the accuracy and validation loss graphs suggested that the models were simply not complex enough to recognise the finer patterns in the images as both losses stayed constant for each model as illustrated by the loss graph of the image augmentation model in Figure 13. This may be the case because the ISIC dataset skin lesion images are very similar to each other and unbalanced as there are a larger proportion of benign images compared to malignant and a smaller model is just incapable of recognising those finer patterns and the model was just classifying every image as being Benign, giving a false evaluation of performance.

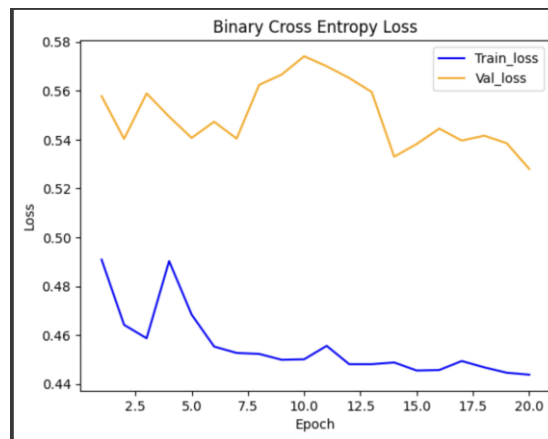


Figure 13. Graph of Cross Entropy Loss against epochs for the Baseline Model with Image Augmentation with orange for validation and blue for training loss.

4.1.2 Transfer Learning Model

These observations suggested the need for deeper and more complex models that may be able to perform better on the dataset. A machine learning technique called Transfer learning was employed to build the next models. Transfer Learning is a technique where a pre-trained model uses its knowledge from one image classification problem to make predictions on a new but similar image classification problem (Castillo, D., 2021). This is less resource-intensive and gives access to more complex models giving it great significance in deep learning.

The first model prepared used the VGG16 model. VGG16, proposed at the University of Oxford in 2014, was one of the top-performing models in the ILSVRC(ImageNet Large Scale Visual Recognition Challenge) competition in 2014 and to date is still widely used as a pre-trained model for image classification tasks. The model has a large and complex architecture consisting of 13 convolutional layers and 3 fully connected layers as depicted in Figure 14.

Model: "model"		
Layer (type)	Output Shape	Param #
input_1 (InputLayer)	[(None, 224, 224, 3)]	0
block1_conv1 (Conv2D)	(None, 224, 224, 64)	1792
block1_conv2 (Conv2D)	(None, 224, 224, 64)	36928
block1_pool (MaxPooling2D)	(None, 112, 112, 64)	0
block2_conv1 (Conv2D)	(None, 112, 112, 128)	73856
block2_conv2 (Conv2D)	(None, 112, 112, 128)	147584
block2_pool (MaxPooling2D)	(None, 56, 56, 128)	0
block3_conv1 (Conv2D)	(None, 56, 56, 256)	295168
block3_conv2 (Conv2D)	(None, 56, 56, 256)	590080
block3_conv3 (Conv2D)	(None, 56, 56, 256)	590080
block3_pool (MaxPooling2D)	(None, 28, 28, 256)	0
block4_conv1 (Conv2D)	(None, 28, 28, 512)	1180160
block4_conv2 (Conv2D)	(None, 28, 28, 512)	2359808
block4_conv3 (Conv2D)	(None, 28, 28, 512)	2359808
block4_pool (MaxPooling2D)	(None, 14, 14, 512)	0
block5_conv1 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv2 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv3 (Conv2D)	(None, 14, 14, 512)	2359808
block5_pool (MaxPooling2D)	(None, 7, 7, 512)	0
flatten (Flatten)	(None, 25088)	0
dense (Dense)	(None, 128)	3211392
dense_1 (Dense)	(None, 2)	258
=====		
Total params: 17,926,338		
Trainable params: 3,211,650		
Non-trainable params: 14,714,688		

Figure 14. Structure of the VGG16 Model.

The classification layers were removed from the VGG16 model and new layers were added to allow the model to make predictions on the new dataset. All the other layers were frozen to preserve the weight information in the model during training. Using the ImageDatagenerator, the mean pixel values were manually set to be the same as that of the ImageNet dataset on which the model was originally trained on. Keeping all other hyperparameters the same as previous models, the model was trained for 10 epochs and an accuracy of 79.461% was achieved. Further evaluation of the loss and accuracy graphs suggested this performance metric may not be the best for this dataset as, even in the case where the model predicts all test images as Benign, it would still achieve a very high score due to the imbalance of classes in the dataset. This major oversight in the preparation of the models prompted the need for the models to be revised.

4.2 SECOND BATCH – AUC

The second batch consisted of revised versions of the first batch with the only change being the switch from Accuracy to AUC performance metric. The results and corresponding Loss and AUC graphs for each model are presented in Table 3. The Baseline model producing an AUC of 65% suggested that the model was performing just a little bit better than just randomly guessing which would be the case if the AUC was 50%. This value, together with the evaluation of the corresponding Loss graph suggested that the model was overfitting and required training with the regularisations added.

As observable in Table 3, the regularisations did the opposite of improving the performance with dropout reducing the performance to as low as 59% and image augmentation to 63%. This decrease in performance may be attributed to the fact that the images in the dataset are very similar to each other and regularisation in conjunction with the small 50x50 size of the images was causing too much of the information in the images to be lost. VGG16, on the hand, showed a 7% improvement in performance but evaluating the loss graph suggested that the model was overfitting and further improvement is achievable if this problem is addressed.

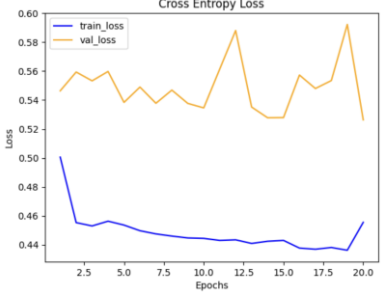
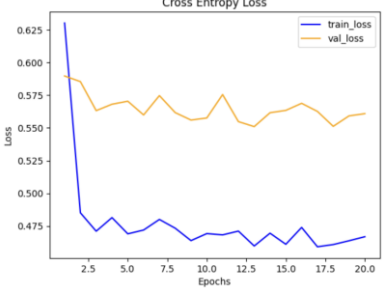
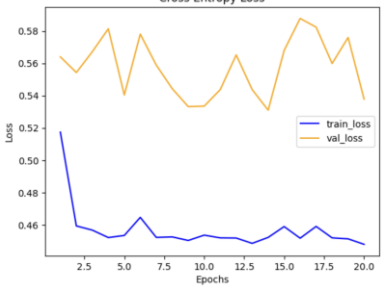
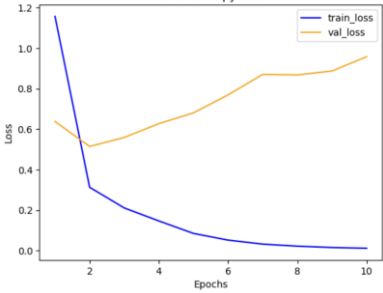
Model	Best AUC in all Epochs	Loss Graph
Baseline	64.29% (Early stopping after 15 epochs)	
Baseline with dropout	60.34% (Early stopping after 18 epochs)	
Baseline with Image Augmentation	62.46% (Early stopping after 13 epochs)	
<u>VGG16 Transfer model</u>	<u>72.65%</u>	

Table 3. Shows results for the different models with the highest values of AUC performance metric during training and loss graphs.

The large difference in the loss values for each graph also suggested that the models were overfitting on the training data and required further changes to improve performance or possibly the number of epochs the model was being trained for simply was not enough.

4.3 THIRD BATCH – LARGER INPUT SIZE

The third batch of models involved switching from the SGD optimiser to the Adam optimiser, which on paper is better for deep learning models. Another major change to the model was the switch from the 50x50 image input size to 224x224 as it was observed that the images being already very similar to each other, the small size was causing too much loss of valuable information in the images, and this was causing the models to not effectively learn from the images. A larger image size would mean longer training times, but this sacrifice was made in exchange for better performance.

In addition to the Baseline model and the VGG16 model, several other models were also trained with the same hyperparameters to test and compare performance as shown in Table 3.

Model	Highest AUC in all Epochs
Baseline	69.31%
VGG16	74.57%
VGG16 with Image Augmentation	67.58%
ResNet50	50.000%
ResNet50 with Image Augmentation	57.73%
<u>VGG19</u>	<u>74.48%</u>
VGG19 with Image Augmentation	64.78%
<u>EfficientNetV2L</u>	<u>78.37%</u>
InceptionV3	56.462%

Table 4. Results for 9 models with 224x224 input size and Adam optimiser.

All models in this batch were trained with 64 batch size for 10 epochs each. Due to consistently dropping performance of models, dropout regularisation was omitted and only Image Augmentation regularisation was used for a few high performing models. The ResNet50 even with regularisation, produced the worst results and was deemed to be incompatible for the dataset. On the other hand, VGG19 and EfficientNetV2L models produced the best results,

74.48% and 78.37% respectively, without adding regularisation. This improved performance may be attributed to the fact that both models have large architectures that are capable of finding very intricate patterns in the images.

The VGG19 model builds upon the earlier VGG16 model by adding 3 more convolutional layers making it deeper and more complex compared to the VGG16 model (Boesch, 2023). The increase in performance from the VGG16 to VGG19 may be attributed to the increase in the number of parameters of the model from 138.4M to 143.7M (Keras, 2023), resulting in the model's ability to learn more complex patterns. On the other hand, the EfficientNetV2L with although 119M parameters, has a much wider and deeper structure compared to other models (Reiff, 2022). This depth and increased width, coupled with its ability to train efficiently on low resources as designed by Google, may be the reason for its much higher performance compared to other models.

4.4 FOURTH BATCH – CATEGORICAL LABELS

Discussing the results of previous batches with the supervisors uncovered an underlying problem that was affecting the performance of the models. This was the unbalanced dataset problem where the benign class due to its larger proportion in the dataset was overshadowing the lesser number of malignant class and causing the model to be specialised in finding patterns in benign images. Two solutions were proposed for the problem and this batch tests the first solution where all the labels were converted from Binary to One Hot encoded labels.

4.4.1 One Hot Encoding

One hot encoding is used to represent categorical data in the form of a binary vector which is easily understood by a machine learning algorithm. In our case, where two classes are involved, one hot encoding turns Benign and Malignant categorical labels to the arrays [1,0] and [0,1] respectively. The intuition behind this is that, when the labels were binary, the model was returning a value ranging between 0 and 1 for predicting a label i.e. it was returning a percentage of any image being malignant. With the imbalance in the datasets, the values prediction probabilities returned by the model with respect to ground truth labels were ranging between 0 and 0.3 as evidenced by Figure 15. That meant that the model was classifying everything under

the 0.5 threshold to be predicted as Benign, resulting in a very one-sided confusion matrix as depicted in Figure 16.

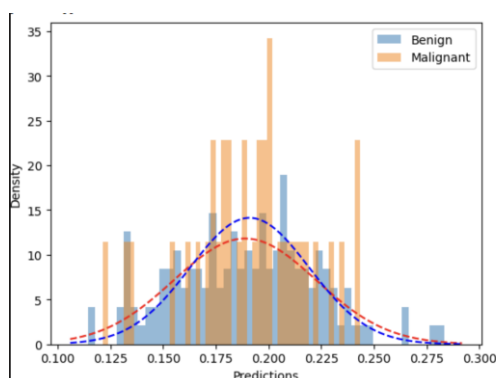


Figure 15. Probability Distribution of predicted labels returned by the model.

Confusion matrix	Predicted Benign	Predicted Malignant
True Benign	483	0
True Malignant	117	0

Figure 16. Confusion matrix for predictions before one hot encoded labels

One hot encoding of the labels was done by setting the class_mode for ImageDataGenerator's flow_from_directory method to categorical and setting the loss function to be categorical cross entropy with a SoftMax layer at the end of the model with 2 output nodes. This method fixed the distribution problem by separating the returned probability in the form as depicted in Figure 17, where first column is the probability for Benign and second for Malignant.

```
[[0.69258505 0.30741498]
 [0.7148171 0.28518295]
 [0.7259521 0.27404785]
 [0.715226 0.28477398]
 [0.6365239 0.363476 ]
 [0.7236082 0.27639177]
 [0.6777174 0.32228267]
 [0.68112046 0.31887954]
 [0.7049012 0.29509884]
 [0.7322603 0.26773974]
 [0.63992566 0.36007437]
 [0.6851693 0.31483075]
 [0.6583476 0.34165242]
 [0.6512582 0.34874183]
 [0.7351859 0.26481408]
 [0.7838587 0.21614122]
 [0.77805907 0.22194096]
 [0.7337423 0.26625767]
 [0.65981585 0.34018412]
 [0.69105786 0.3089422 ]
 [0.7039284 0.29607168]
 [0.62640667 0.37359336]]
```

Figure 17. Probabilities returned by mode, first column for Benign and second for Malignant.

Using argmax to return the index for of the larger value in each row returned a 0 for the first column and a 1 for second, conveniently in the correct form for us to create a confusion matrix where 0 represents Benign and 1 represents Malignant.

4.4.2 Models and Results for One Hot encoded

Models	AUC	Confusion Matrix			Correct Malignant Classification
Baseline	57.30%	Confusion matrix	Predicted Benign	Predicted Malignant	26%
		True Benign	319	164	
		True Malignant	86	31	
Baseline with Dropout	61.40%	Confusion matrix	Predicted Benign	Predicted Malignant	65%
		True Benign	157	326	
		True Malignant	41	76	
VGG16	71.03%	Confusion matrix	Predicted Benign	Predicted Malignant	14%
		True Benign	405	78	
		True Malignant	101	16	
VGG19	72.12%	Confusion matrix	Predicted Benign	Predicted Malignant	24%
		True Benign	368	115	
		True Malignant	89	28	

Table 5. AUC results and confusion matrices from fourth batch of models.

The results as illustrated in Table 5, on paper looked good but after further evaluation of the confusion matrices produced, the technique was producing very inconsistent results. Baseline model produced an alright confusion matrix where 65% of the malignant images were correctly

classified but on the other hand 51% of the benign images were incorrectly classified. Model with dropout produced a 99% malignant classification accuracy but at the same time classified all benign images as malignant as well which is unacceptable. VGG16 in contrast had a slight improvement from previous implementation with one hot encoded label, producing 29% correct malignant classifications and an 8% decrease in incorrect benign classifications.

These inconsistencies may again be attributed to the small size of the dataset where even after applying weighting to each class, the models still required either more data or more training time to find meaningful improvements in the performance. This inconsistency in the results deemed this method ineffective use for further models.

4.5 FIFTH BATCH – SMALLER BATCH SIZE

The fifth batch of models tested the effect of having a smaller batch size on the performance. In theory, a smaller batch size means that a model can update its parameters more frequently, converging faster to reach an optimal performance. But this also means that training times increase as more iterations are required to reach convergence.

Models	Highest AUC in all Epochs	Confusion Matrix									
VGG19 with Binary Labels	66.43%	<table> <tr> <th>Confusion matrix</th><th>Predicted Benign</th><th>Predicted Malignant</th></tr> <tr> <th>True Benign</th><td>483</td><td>0</td></tr> <tr> <th>True Malignant</th><td>117</td><td>0</td></tr> </table>	Confusion matrix	Predicted Benign	Predicted Malignant	True Benign	483	0	True Malignant	117	0
Confusion matrix	Predicted Benign	Predicted Malignant									
True Benign	483	0									
True Malignant	117	0									
VGG19 with one hot encoding	68.56%	<table> <tr> <th>Confusion matrix</th><th>Predicted Benign</th><th>Predicted Malignant</th></tr> <tr> <th>True Benign</th><td>441</td><td>42</td></tr> <tr> <th>True Malignant</th><td>90</td><td>27</td></tr> </table>	Confusion matrix	Predicted Benign	Predicted Malignant	True Benign	441	42	True Malignant	90	27
Confusion matrix	Predicted Benign	Predicted Malignant									
True Benign	441	42									
True Malignant	90	27									
EfficientNetV2L with binary labels	77.42%	<table> <tr> <th>Confusion matrix</th><th>Predicted Benign</th><th>Predicted Malignant</th></tr> <tr> <th>True Benign</th><td>483</td><td>0</td></tr> <tr> <th>True Malignant</th><td>117</td><td>0</td></tr> </table>	Confusion matrix	Predicted Benign	Predicted Malignant	True Benign	483	0	True Malignant	117	0
Confusion matrix	Predicted Benign	Predicted Malignant									
True Benign	483	0									
True Malignant	117	0									

EfficientNetV2L with one hot encoding labels	77.28%	<table> <tr> <th>Confusion matrix</th><th>Predicted Benign</th><th>Predicted Malignant</th></tr> <tr> <th>True Benign</th><td>417</td><td>66</td></tr> <tr> <th>True Malignant</th><td>85</td><td>32</td></tr> </table>	Confusion matrix	Predicted Benign	Predicted Malignant	True Benign	417	66	True Malignant	85	32
Confusion matrix	Predicted Benign	Predicted Malignant									
True Benign	417	66									
True Malignant	85	32									
EfficientNetV2L with one hot encoding and Image Augmentation	51.17%	<table> <tr> <th>Confusion matrix</th><th>Predicted Benign</th><th>Predicted Malignant</th></tr> <tr> <th>True Benign</th><td>318</td><td>165</td></tr> <tr> <th>True Malignant</th><td>72</td><td>45</td></tr> </table>	Confusion matrix	Predicted Benign	Predicted Malignant	True Benign	318	165	True Malignant	72	45
Confusion matrix	Predicted Benign	Predicted Malignant									
True Benign	318	165									
True Malignant	72	45									

Table 6. AUC and Confusion matrices for fifth batch of models.

As visible in Table 6, the results were not extraordinary and generally had a decline in performance from previous models, except for the EfficientNetV2L models that continued to perform well on the dataset. The lack of an expected improvement in performance may be attributed to the fact that although a smaller batch size may improve performance by allowing the model to generalise better and converge faster, it may also introduce unwanted noise in the training process that could prove detrimental to the training process. This noise may also be the reason for the sharp decline in performance for the EfficientNetV2L model with image augmentation as augmenting images also adds noise to the dataset and the resulting compounded noise meant that the model degraded to random guessing. Overall, the test proved that a batch size of 64 was adequate for the training process.

4.6 CONCLUSION

In conclusion, this chapter highlighted the progress made in developing and refining models for single image classification of skin lesions. These findings emphasised the importance of appropriate evaluation metrics, data pre-processing techniques, and model configurations. It is through iterative experimentation and careful refinement that more effective models can be developed, ultimately leading to better results.

5 UGLY DUCKLING

Taking inspiration from the research carried out by Mohseni et al. in the paper “Can self-training identify suspicious ugly duckling lesions?” as discussed previously in the Ugly Duckling chapter, this was an attempt at implementing an anomaly detection technique using auto encoders.

5.1 AUTOENCODER

An Autoencoder is a type of feed-forward artificial neural network that comprises three main parts, an encoder, a bottleneck layer and a decoder as depicted in Figure 18. The encoder layers receive input data and gradually compress it into a lower dimensional representation by consistently decreasing layer sizes. The bottleneck layer, located at the final stage of the encoder, plays a vital role in the autoencoder by effectively reducing the data's dimensionality while preserving essential information. This enables models to effectively filter out the noise and focus solely on capturing the key elements within the data. The size of this layer determines how much compression and dimensionality reduction is done where a small bottleneck layer allows for more compression but produces a lower quality reconstruction due to loss of information and conversely, a larger bottleneck layer retains more information, producing a higher quality output but may lead to overfitting. This lower dimensional representation of data is called a Latent-Space Representation and will be referred to as such in the following paragraphs. Finally, the decoder layers take this latent-space representation, increasing the dimensionality back to the original dimensions by consistently increasing layer sizes, and try to reconstruct the input data as closely as possible. Any differences in this reconstruction are referred to as reconstruction errors. (Dertat, 2017)

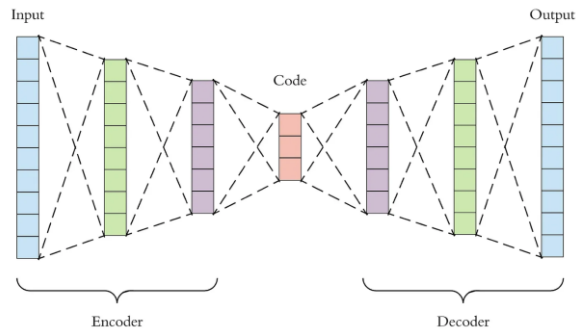


Figure 18. Structure of an Autoencoder (Source: Dertat, 2017:online)

The intuition behind using an auto encoder for anomaly detection tasks is that when an auto encoder is trained on a class of images, Benign images in our case, the model learns to identify the most significant features of Benign images and this information is stored in the bottleneck layer as a latent-space representation. If during testing, an image of a malignant skin lesion is passed to the model, the model would not find the same features as the benign images, resulting in a different latent-space representation, so the reconstruction error of the output produced from this compared to benign images would be much higher, suggesting that the image may be anomalous.

Another way of identifying an anomaly is to calculate the Probability Density Function of the latent-space representation produced after training on the normal data, using a technique called Kernel Density Estimation (KDE). KDE is technique where a kernel is placed at every data point and then the contributions of these kernels at each point are summed up to create a smooth, continuous curve that represents an estimate of the underlying PDF (Lets Talk About Science, 2020). As illustrated in Figure 19, Gaussian kernels are placed at each data point and summed up to create a graph of the dataset's underlying probability distribution. When a new input data is passed to the model, the model first creates the latent-space representation for this new input data, calculates its Probability Density Function using the same KDE algorithm as the training data and then depending on how different this PDF is compared to the normal data's PDF, it can be determined if the data may be anomalous or not.

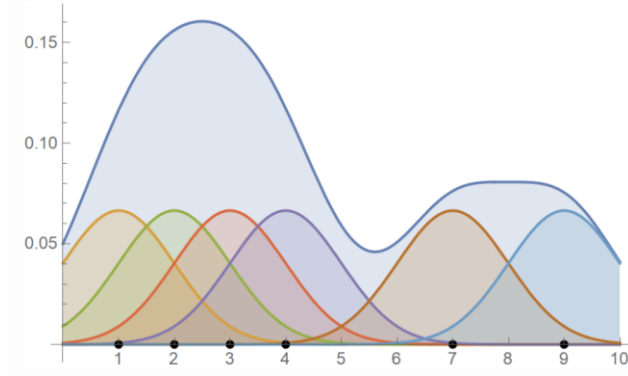


Figure 19. Individual Gaussian kernels plotted along with the Kernel Density Estimate. (Source: Lets talk About Science, 2020:online)

5.2 PREPARATION

An auto encoder with four convolution layers in the encoding and decoding halves was created with the first layer having a size 128 which was decreased to 16 in consequent layers up to the bottleneck. The encoder half of the model employed several max pooling layers to reduce the dimensionality of the data and conversely, the decoder half was equipped with up sampling layers to gradually increase the dimensions back to the same dimensions as the input data. Finally, an output convolutional layer with a sigmoid function and 3 output nodes was added to generate an RGB image.

5.3 DATASET

The Data used to train the model came from the original three datasets that were used in the earlier models. Additionally, corresponding submask images for each dataset were also downloaded from the official ISIC website. These binary submask images were used to generate segmentations of the normal image containing only the skin lesions, as depicted in Figure 20. The Datasets were then rearranged, moving all segmented benign images from the test and train dataset to create new training dataset. The malignant images from the training, test and validation datasets were all moved to a new test dataset. The benign images inside the validation dataset were not moved and used for validation during training. The final dataset breakdown is detailed below in Table 7.

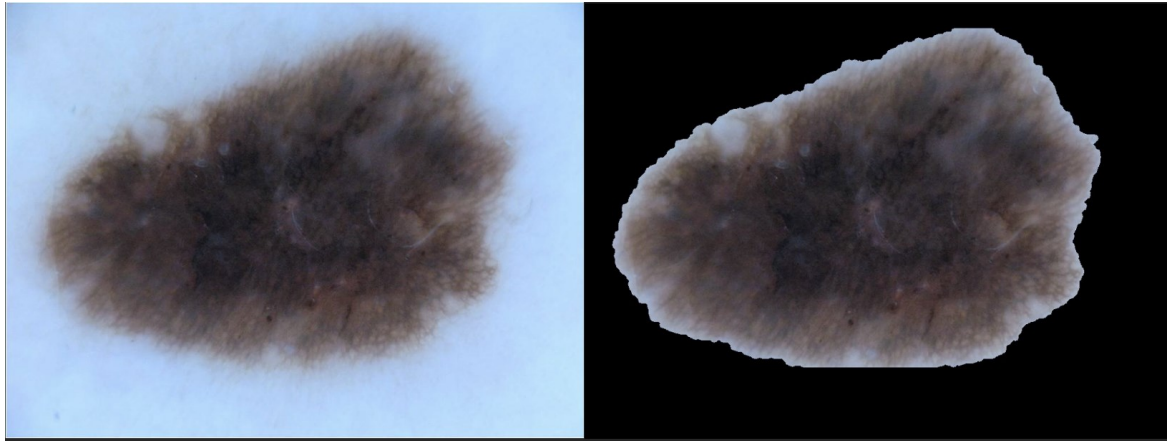


Figure 20. Lesion image before and after segmentation

Dataset	Images
Train	2109 Benign Images
Validation	120 Benign Images
Test	521 Malignant Images

Table 7. Breakdown of the datasets to be used for Training and Testing of the Autoencoder.

5.4 TRAINING

The model was trained with Mean Squared Error (MSE) as the loss function and performance metric for 20 epochs. MSE allowed the model to calculate the reconstruction error between the input and output images and using backpropagation, to get as good an output as possible. Figure 21 shows some inputs and corresponding output images for the model.

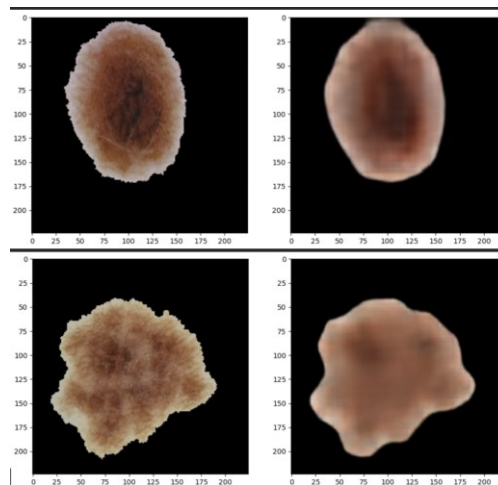


Figure 21. Input images and corresponding outputs

As observable in the Figure 21, the outputs although not great quality, were acceptable for a model trained for only 20 epochs on a small dataset of 2109 images. Deeming this model good enough at reconstructing the benign images, we moved on to the anomaly detection stage.

5.5 ANOMALY DETECTION

The encoder half of the model, with the preserved weights, was prepared to carry out encoding on images to be tested. The latent-space representation produced from the prediction method of the encoder for the benign images was used to fit a KDE. This KDE was then used to get a density score for new images to be tested for an anomaly. At the same time, the new images were also passed to the prediction method of the original autoencoder to produce reconstruction errors for the images. Comparing density and reconstruction errors in conjunction with their respective standard deviations, threshold values were determined to classify images as normal or anomalous. Any image with a density lower than the threshold, or any image with a reconstruction error higher than the threshold was classified as anomalies.

The original test dataset with 483 Benign and 117 Malignant images was segmented and used to test the model's performance. Predictions were made for these images and a confusion matrix was produced with the aid of the groundtruth labels. The model results were average as evident in the confusion matrix in Figure 22.

Confusion matrix	Predicted Benign	Predicted Malignant
True Benign	225	258
True Malignant	48	69

Figure 22. Confusion matrix for the predictions

58% of the malignant images were correctly identified as anomalies but at the same time, 53% of the Benign images were also flagged as anomalies. Further evaluation of the density and reconstruction errors and their respective standard deviations revealed that the values produced were not as they were expected. The standard deviation for the density of the benign images was nearly zero and the reconstruction error and standard deviation compared to malignant images were much higher as evident in Figure 23. These contradicting expected values produced by the

model may be attributed to the fact that the model was trained on a relatively very small dataset for only 20 epochs as evidenced by the loss graph in Figure 24, the model would have to be trained for 100s of epochs to have a chance of finding any meaningful improvements in performance.

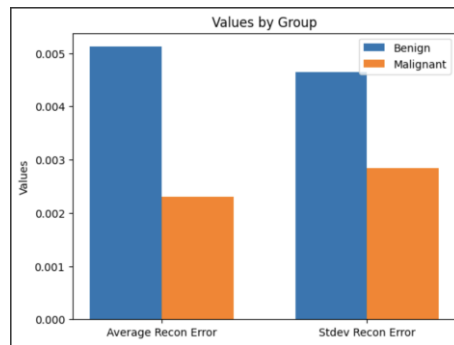


Figure 23. Reconstruction errors and standard deviations for Benign and Malignant Images

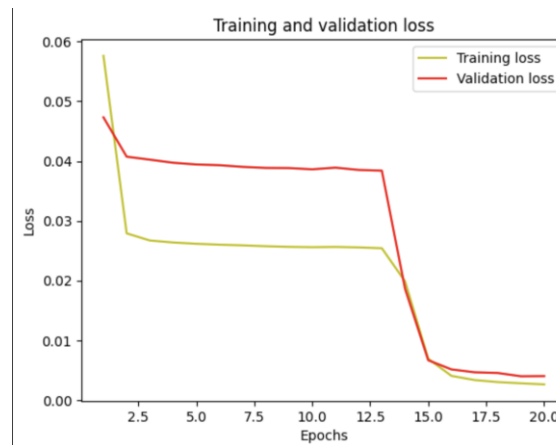


Figure 24. MSE loss graph for the model.

5.6 CONCLUSION

In conclusion, the utilization of autoencoders for anomaly detection presents both opportunities and challenges. Throughout this chapter, the application of autoencoders in detecting anomalies was explored and the main problem encountered in training was discussed. The primary issue encountered was the need for a large number of training epochs before notable performance improvements could be observed. Autoencoders require substantial amounts of data to

effectively learn the underlying patterns and capture the complexity of the normal data distribution. With the limited data, it struggled to reach its full potential in anomaly detection.

6 EVALUATION

This section aims to evaluate the extent to which the predefined aims and objectives of the project have been fulfilled. This will be accomplished through a critical analysis of the project's work, tests conducted, and the obtained results.

6.1 SINGLE IMAGE CLASSIFICATION

The first aim of the project to analyse the use of NNs in skin lesion classification was done by first carrying out a thorough literature review of the currently available work and using this knowledge to produce models that classify skin cancer lesions. The Models and Results section of the report walked through the process of iteratively making changes to the models to achieve a gain in performance.

The first batch of models that were trained with binary labels, 64 batch size, SGD optimiser and Accuracy as the performance metric, produced results ranging from 76% to 78%, that on paper appeared to be very good as a starting point. However, a critical oversight was identified in the setup of the first batch of models. It was discovered that the dataset having four times as many benign images as malignant, ones resulted in a misleading sense of high performance, as if it was the case that all images were classified as benign, the model would still produce an accuracy of 75% which does not accurately reflect the model's ability to distinguish between the two classes.

The second batch of models that were trained with AUC as the performance metric, produced a realistic measure of performance with results ranging between 60% to 72%. This was lower than the values resulting from the Accuracy performance metric because AUC gives a more accurate measure of performance by giving the true positive rate of predictions against the false positive rate of predictions. This is an overall measure of the model's performance irrespective of the class distribution in the dataset.

Third batch models with ADAM as the optimiser and 224x224 image sizes produced considerably better results with high values ranging between 74% and 78%. This was the result of increasing the image sizes which retain more information than the smaller images, allowing

models to find more intricate patterns and features in the training images. EfficientNetV2L with its wider and more complex structure and ability to train efficiently on lower resources produced the best results. Although, at the same time, models such as the ResNet50 and InceptionV3 that were discussed in the literature review, produced the worst results, ranging between 50% and 60%. A possible cause for this may be that these models require a greater number of epochs to learn meaningful features in an image and the accessibility to good hardware resources for the researchers discussed about in the literature review could train these models for longer periods, resulting in better results.

Fourth and fifth batch of models testing the models with one hot encoded labels and then reducing the batch size for these models produced similar levels of results as previous models but much better confusion matrices as one hot encoding was returning probabilities for images belonging to either class and not just being in the range of 0 and 0.3 as was the case with binary labels.

After evaluating these results and understanding the limitations that revolve around skin cancer lesion classification, I believe I have met the first aim of the project.

6.2 UGLY DUCKLING AND DR ABCDE

The aim to investigate the use of these methods in Neural Network was met as after extensive literature review and then implementing a starting point for an anomaly detection system using auto encoders is a good first step in the direction of solving this major challenge in the field of machine learning.

The anomaly detection system produced subpar results with reconstruction error values for the benign class being much higher than the malignant class. This may be due to the lack of accessibility to high end resources to run the model for a greater number of epochs. It is expected that if the dataset is also grown in size and model is allowed to train for longer, a possible performance gain is achievable.

7 CONCLUSION AND REFLECTION

In conclusion, the experience of carrying out this large research project has been both challenging and rewarding. One of the significant aspects of this journey was diving into the realm of deep learning, an area that I had no prior experience with. Through this project, I was able to gain valuable knowledge and skills in deep learning techniques, including neural networks, convolutional neural networks (CNNs), and their applications in skin cancer detection. This newfound expertise opens up a plethora of possibilities for future research and contributions to the field.

Another important lesson I learned throughout this project is the paramount importance of data preparation. While the models and algorithms play a crucial role, it became evident that the quality and suitability of the data have a profound impact on the performance and outcomes of the system. The effort invested trying to reduce effects of the imbalance in the dataset proved to be instrumental in enhancing the performance and reliability of the models. This experience emphasized the notion that data is the foundation upon which successful deep learning models are built.

Carrying out this research project has not only contributed to advancing my understanding of deep learning but has also instilled in me a sense of accomplishment and the desire to continue exploring the intersection of artificial intelligence and healthcare. The challenges encountered throughout the project have served as valuable learning experiences and have further motivated me to pursue further research in the field of medical image analysis and anomaly detection.

7.1 FUTURE WORK

In the future, there are several avenues for further work and improvements based on the findings and limitations of the current research project. The following points can be considered for future work:

The results of the prepared models can be further improved by having access to high-end computational resources. With more powerful hardware, larger datasets can be utilized for training, allowing the models to capture a wider range of patterns and improve their performance.

Increased computational resources would also enable training the models for a higher number of epochs, potentially leading to better convergence and overall performance. This would help address the limitations observed regarding the need for an extensive training period to achieve significant improvements on small datasets.

The anomaly detection system developed in this research project can be further enhanced to further leverage the power of autoencoders. Implementing advanced techniques, such as ensembling multiple models or utilizing transfer learning from pre-trained models, can also be explored to boost the system's performance and robustness.

Moreover, investigating VAEs (Variational Autoencoders) as reviewed in the literature review, could provide insights into their suitability for skin cancer detection. As highlighted in the literature review, variational autoencoders (VAEs) offer potential advantages in capturing latent representations and generating more diverse samples. Future work can focus on implementing VAEs within the skin cancer detection system and evaluating their performance in comparison to traditional autoencoders.

8 REFERENCES

- [1] 1.5. *Stochastic Gradient Descent* (n.d.) scikit-learn. [Online] [Accessed on 4 May 2023] <https://scikit-learn.org/stable/modules/sgd.html>.
- [2] *A Gentle Introduction to the Adam Optimization Algorithm for Deep Learning* (n.d.) Machinelearningmastery.com. [Online] [Accessed on 4 May 2023] <https://machinelearningmastery.com/adam-optimization-algorithm-for-deep-learning/>.
- [3] *A Gentle Introduction to the Rectified Linear Unit (ReLU)* (n.d.) Machinelearningmastery.com. [Online] [Accessed on 10 April 2023] <https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/>.
- [4] An, J. and Cho, S. (n.d.) *Variational Autoencoder based Anomaly Detection using Reconstruction Probability*. Snu.ac.kr. [Online] [Accessed on 17 March 2023] <http://dm.snu.ac.kr/static/docs/TR/SNUDM-TR-2015-03.pdf>.
- [5] Barata, C. (n.d.) *ISIC Skin Image Analysis Workshop @ ECCV 2022*. Isic-archive.com. [Online] [Accessed on 13 March 2023] <https://workshop2022.isic-archive.com/>.
- [6] Bi, L., Kim, J., Ahn, E. and Feng, D. (n.d.) *Automatic skin lesion analysis using large-scale Dermoscopy images and deep residual networks*. Arxiv.org. [Online] [Accessed on 17 March 2023] <https://arxiv.org/pdf/1703.04197.pdf>.
- [7] Bisong, E. (2019) 'What is deep learning?' *In Building Machine Learning and Deep Learning Models on Google Cloud Platform*. Berkeley, CA: Apress, pp. 327–329.
- [8] Boesch, G. (2021) *VGG Very Deep Convolutional Networks (VGGNet) - What you need to know*. viso.ai. [Online] [Accessed on 9 May 2023] <https://viso.ai/deep-learning/vgg-very-deep-convolutional-networks/>.
- [9] Brenner, M. and Hearing, V. J. (2008) 'The protective role of melanin against UV damage in human skin.' *Photochemistry and photobiology*, 84(3) pp. 539–549.
- [10] Brinker, T. J., Hekler, A., Utikal, J. S., Grabe, N., Schadendorf, D., Klode, J., Berking, C., Steeb, T., Enk, A. H. and von Kalle, C. (2018) 'Skin cancer classification using convolutional neural networks: Systematic review.' *Journal of medical internet research*, 20(10) p. e11936.

- [11] Cassidy, B., Kendrick, C., Brodzicki, A., Jaworek-Korjakowska, J. and Yap, M. H. (2022) 'Analysis of the ISIC image datasets: Usage, benchmarks and recommendations.' *Medical image analysis*, 75(102305) p. 102305.

- [12] *Classification: ROC curve and AUC* (n.d.) Google for Developers. [Online] [Accessed on 5 May 2023] <https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc>.

- [13] Dertat, A. (2017) *Applied Deep Learning - Part 3: Autoencoders*. Towards Data Science. [Online] [Accessed on 13 May 2023] <https://towardsdatascience.com/applied-deep-learning-part-3-autoencoders-1c083af4d798>.

- [14] ETH Zurich (2010) 'Molecular switch prevents cell damage caused by UVB rays from sunlight.' *Science Daily*.

- [15] Foote, K. D. (2022) *A brief history of deep learning*. DATAVERSITY. [Online] [Accessed on 2 April 2023] <https://www.dataversity.net/brief-history-deep-learning/>.

- [16] Ganesh, P. (2020) *From LeNet to EfficientNet: The evolution of CNNs*. Towards Data Science. [Online] [Accessed on 5 April 2023] <https://towardsdatascience.com/from-lenet-to-efficientnet-the-evolution-of-cnns-3a57eb34672f>.

- [17] Grove, R. and Green, R. (2020) 'Melanoma and nevi classification using convolution neural networks.' In *2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ)*. IEEE, pp. 1–6.

- [18] *How Do Convolutional Layers Work in Deep Learning Neural Networks?* (n.d.) Machinelearningmastery.com. [Online] [Accessed on 6 April 2023] <https://machinelearningmastery.com/convolutional-layers-for-deep-learning-neural-networks/>.

- [19] *How to Classify Photos of Dogs and Cats (with 97% accuracy)* (n.d.) Machinelearningmastery.com. [Online] [Accessed on 25 April 2023] <https://machinelearningmastery.com/how-to-develop-a-convolutional-neural-network-to-classify-photos-of-dogs-and-cats/>.

- [20] *ISIC Archive* (n.d.) Isic-archive.com. [Online] [Accessed on 13 March 2023]
<https://www.isic-archive.com/#!/topWithHeader/tightContentTop/about/aboutIsicOverview>.

- [21] Jojoa Acosta, M. F., Caballero Tovar, L. Y., Garcia-Zapirain, M. B. and Percybrooks, W. S. (2021) 'Melanoma diagnosis using deep learning techniques on dermoscopic images.' *BMC medical imaging*, 21(1) p. 6.

- [22] Kamperis, S. (2020) 'A gentle introduction to kernel density estimation.' Let's talk about science! 8th December. [Online] [Accessed on 14 May 2023]
<https://ekamperi.github.io/math/2020/12/08/kernel-density-estimation.html>.

- [23] *Keras applications* (n.d.) Keras.io. [Online] [Accessed on 9 May 2023]
<https://keras.io/api/applications/>.

- [24] *Looking at your mole or skin (dermoscopy)* (n.d.) Cancerresearchuk.org. [Online] [Accessed on 1 March 2023] <https://www.cancerresearchuk.org/about-cancer/melanoma/getting-diagnosed/tests-diagnose/looking-your-mole-dermoscopy>.

- [25] Matt (n.d.) *Fitzpatrick17k*.

- [26] *Melanoma skin cancer statistics* (2015) Cancer Research UK. [Online] [Accessed on 1 March 2023] <https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/melanoma-skin-cancer>.

- [27] *Melanoma warning signs and images* (2019) The Skin Cancer Foundation. [Online] [Accessed on 1 March 2023] <https://www.skincancer.org/skin-cancer-information/melanoma/melanoma-warning-signs-and-images/>.

- [28] Mohseni, M., Yap, J., Yolland, W., Koochek, A. and Atkins, M. S. (2021) 'Can self-training identify suspicious ugly duckling lesions?' *arXiv [cs.CV]*.

- [29] Reiff, D. (2022) *Generalizing your model: An example with EfficientNetV2 and cats & dogs*. Towards Data Science. [Online] [Accessed on 9 May 2023]
<https://towardsdatascience.com/generalizing-your-model-an-example-with-efficientnetv2-and-cats-dogs-6903740dfe2c>.

- [30] *resnet50* (n.d.) Mathworks.com. [Online] [Accessed on 16 March 2023]
<https://www.mathworks.com/help/deeplearning/ref/resnet50.html>.

- [31] Saha, S. (2018) *A comprehensive guide to convolutional neural networks — the ELI5 way*. Towards Data Science. [Online] [Accessed on 9 April 2023] <https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53>.
- [32] *Stages of melanoma* (2019) AIM at Melanoma Foundation. [Online] [Accessed on 1 March 2023] <https://www.aimatmelanoma.org/stages-of-melanoma/>.
- [33] *Tests for melanoma skin cancer* (n.d.) Cancer.org. [Online] [Accessed on 10 March 2023] <https://www.cancer.org/cancer/melanoma-skin-cancer/detection-diagnosis-staging/how-diagnosed.html>.
- [34] Verma, Y. (2021) *A complete understanding of dense layers in neural networks*. Analytics India Magazine. [Online] [Accessed on 10 April 2023] <https://analyticsindiamag.com/a-complete-understanding-of-dense-layers-in-neural-networks/>.
- [35] *What is melanoma?* (n.d.) Cancerresearchuk.org. [Online] [Accessed on 14 February 2023] <https://www.cancerresearchuk.org/about-cancer/melanoma/about>.
- [36] *What to look for: ABCDEs of melanoma* (n.d.) Aad.org. [Online] [Accessed on 1 March 2023] <https://www.aad.org/public/diseases/skin-cancer/find/at-risk/abcdes>.
- [37] Yegulalp, S. (2022) *What is TensorFlow? The machine learning library explained*. InfoWorld. [Online] [Accessed on 24 April 2023] <https://www.infoworld.com/article/3278008/what-is-tensorflow-the-machine-learning-library-explained.html>.

CAN AI SPOT SKIN CANCER LESIONS

BY: MUHAMMAD ALI SYED

WHY SKIN CANCER

- Skin Cancer Curable if diagnosed early
- Dermatologists unreliable
- Neural Networks better option

AIMS AND OBJECTIVES

Aims:

Analyse the use of Neural Networks in skin cancer detection and research the implementation of the Ugly Duckling and DR ABCDE methods for Neural Networks.

Objectives:

- Research and test current skin cancer detection methods
- Investigate the Ugly Duckling method
- Investigate the 7-point checklist for the DR ABCDE method
- Analyse the performance and evaluate findings in comparison to previous research

LIT REVIEW

- **Deep Learning**
 - Theory and concepts
- **Normal Image Classification**
 - Data preparation
 - Baseline models
- **Ugly Duckling**
 - Auto Encoders
- **DR ABCDE**

DEVELOPMENT

- **Notebooks**
- **Datasets**
- **Earlier Models**
- **Later Models**

EVALUATION

- Base Models
- EfficientNetV2L
- VGG19
- Other Transfer Models
- Auto Encoder

EfficientNetV2L	78.37%
EfficientNetV2L with binary/one hot encoded labels and 32 batchsize	77.3%
VGG19	74.48%
VGG16	74.57%

FUTURE WORK

- Bigger Datasets
- Auto encoders for Ugly Duckling

CONCLUSION

- Research
- Plan
- Met Aims?

Video Demonstration: [Video Demonstration.mp4](#)