Skin Cancer Pattern Extraction and Prediction Using Deep Convolutional Neural Networks

ABSTRACT

One of the most common cancers in Australia is skin cancer. It is one of the highest rates of skin cancer globally. It is predicted that two out of three Australians would be diagnosed with skin cancer in their lifetime. There are three types of skin cancers: melanomas, basal cell carcinoma (BCC) and squamous cell carcinoma. These cancers require immediate treatment in order to increase the chances of survival. However, the current diagnosis method is time consuming where it involves meetings with skin specialists and medical imaging. Traditional skin cancer early identification methods rely on certain characteristics of lesions. For example, whether the mole's border is defined, asymmetric, its colour, size and diameter and whether the mole evolves. With vast new technology becoming available, it is now possible to pull these features from an image and classify a lesion as either malignant (i.e, dangerous) or benign (i.e., not dangerous) by using neural network methods. As of now, researchers have been using 2D images to detect features of malignant lesions using the neural network. Cost is an issue for both potential patients and the medical industry. With today's medical advancement, skin cancer can be cured as long as diagnosis is done in a timely manner. Therefore, providing a method that could assist in the diagnosis of skin cancer cost-effectively and in a timely manner is a great priority to society.

Our proposed model, using Convolutional Neural Networks, will allow potential patients to assess their risk of having malignant lesions based on a clinical image. This would streamline the diagnostic process by offering a quick, effective and cheap solution. We used five pretrained models, ResNet50, InceptionV3, Xception, VGG-16 and DenseNet.

Our final model ResNet50 produced the best validation accuracy of almost 90% on HAM10000 data. However, putting more weight towards the cancerous classes 'mel' and 'bcc' allowed the model to improve recall scores from an average of 67% to an average of 82%. Test performance of custom weighted ResNet50 trained on ISIC-2019 dataset resulted in 78% validation accuracy. Test performance of custom weighted ResNet50 trained on PAD-UFES-20 dataset resulted in 74% validation

accuracy, however 0% recall scores for the melanoma class. Our highest accuracy of 90% was achieved using resolution 512x512 while the original dimension was 600x450. Although higher resolution improves the model accuracy, it slows the model training time. As a result, our maximum number of epochs was 33 with one available GPU.

Due to time constraints and lack of 3D and Sequential Digital Dermoscopy Imaging (SDDI) datasets, future studies should look into 3D and SDDI skin image analysis and the potential of having a combination of 3D classification model with 2D image analysis. We were restricted to implement ensembling methods due to computational restrictions, limited memory and GPU allocation. While the preprocessing methods such as dull razor and soft attention mapping utilized in the project ultimately were not used for the final models, implementations to perform these processes were retained as they may potentially produce better results with further exploration.

Table of Contents

1. INTRODUCTION

Skin cancer is a common cancer primarily caused by overexposure to ultraviolet radiation (UVR), altering the DNA structure of skin cells (Australian Institute of Health and Welfare [AIHW], 2015). Australia has the highest rate of skin cancer occurrence in the world (Bray et al., 2018). It is predicted that two in three Australians will be diagnosed with a form of skin cancer before they reach 70 years old (Staples et al., 2002). The economic burden of the disease in 2010 in NSW alone is estimated to be $536 million (Doran et al., 2015). Consequently, the difficulties skin cancer poses on individuals diagnosed with the disease, as well as individuals who are indirectly affected are also immense, with additional financial, emotional, and physical strain and impaired wellbeing. If detected early however, skin cancers like melanoma can be treated effectively with a high chance of survival (AIHW, 2015).

Skin cancer detection diagnosis is usually performed by a dermatologist who visually examines the skin area in question. Although this is the typical first step in diagnosis, it was found that dermatologists only correctly diagnosed or classified the early stages of melanomas with an accuracy of 65% - 85% using a hand held microscope (i.e., dermoscopy) (Argenziano & Soyer, 2001). With the advancement of technology, however, neural networks and machine learning models are gradually being utilised to solve similar image recognition and classification problems. Specifically, convolutional neural networks (CNN) are considered the gold standard due to its ability to accurately detect important features in images and thus classify images accurately (Sharma et al., 2018).

Examples include detection rates of 99.3% with the CIFAR-10 dataset (Kolesnikov et al., 2020) and 96.91% with the Fashion-MNIST dataset (Tanveer et al., 2020). For melanoma detection (not using HAM10000 data), examples include 92% accuracy using the CNN ResNet (Gouda & Amudha, 2020) and 96% accuracy using the CNN AlexNet (Hosny et al., 2019). Within the context of skin cancer classification, most CNNs are trained on 2-dimensional (2D) images of skin lesions. Our approach to image analysis is novel in nature as most research on skin lesion classification aims to detect each class equally. As melanoma and basal cell carcinoma skin cancers are identified as the most important and dangerous types of skin lesions, it is desirable to

retain these lesions as their own class, and merge non-dangerous classes, totalling three classes (‘melanoma’, ‘basal cell carcinoma’ and ‘other lesions’). In addition to the three-class identifier, if the class of the lesion was identified as not cancerous (i.e., ‘other lesions’ class), then a second seven-class classification model identifies what the lesion is called as well as produces the probability of that identification against the other classes. Our aim is to use the mentioned approach by comparing and evaluating the performances of five well known pre-trained CNN models: ResNet (He et al., 2016), InceptionV3 (Szegedy et al.,2014), Xception (Chollet, F. (2017), VGG-16 (Simonyan & Zisserman, 2014) and DenseNet (Huang et al., 2016).

The dataset chosen for this comparative analysis is titled Human Against Machine with 10,000 training images (HAM10000) (Tschandl et al., 2018). It contains 10,015 450x600x3 coloured high-resolution images of skin lesions labelled into seven classes. The HAM10000 dataset is commonly used as a skin cancer image classification benchmark for machine learning models, including CNNs (Tschandl et al., 2018). The authors of this paper trained the five CNNs: ResNet, InceptionV3, Xception, VGG-16 and DenseNet with hyper-parameters via the HAM10000 dataset, and detailed the performance of these models. We also tested the final model ResNet50 with custom weights on two external datasets being ISIC-2019 (Codella et al., 2017; Combalia et al., 2019; Taschandl et al., 2018) PAD-UFES-20 (Pacheco et al., 2020). The benefits of this study include a novel approach to melanoma detection using CNNs, as well as an important aid for melanoma diagnosis in clinical settings.

2. Literature Review

In the literature, most successful techniques utilised on skin lesion images are completed by comparing different CNNs against each other (Saba et al., 2019; Satheesha et al., 2017; Tschandl et al., 2018). This involves utilising datasets which are in the form of 2D images. It is important to understand how studies effectively use different pre-trained CNN models for a better outcome in the proposed project. We therefore conduct a literature review analysing such modalities. The literature is sought on the University of Sydney library database, as well as Google Scholar for peer-reviewed journal articles. Keywords used for the searches include: “2D”, “skin cancer”, “CNN”, “melanoma” , “skin lesion”, “skin classification models” and “image analysis”. As machine learning and similar techniques are fairly new

innovations, the date specified for such searches range from 2006 to the present. In this review, a brief background into skin cancer is introduced and secondly 2D image analysis is discussed.

Brief Background

There are three types of skin cancer: 1) melanoma, 2) basal cell carcinoma (BCC) and 3) squamous cell carcinoma (SCC). For the current study and dataset used, we will be focusing on the first two skin cancers. Overexposure to UVR primarily from the sun is one of the main risk factors of skin cancer formation (AIHW, 2015). Additional risk factors include those with fair skin complexion that burn and freckle easily, presence of many moles (i.e., 20 or greater), family or personal history of melanoma, being a male and those aged 50 years old or older (Cancer Institute NSW, 2015). Of the three skin cancer types, melanoma is the most dangerous form of skin cancer as it can rapidly spread to other organs of the body (AIHW, 2015). It is also the most common type of cancer faced by young Australians (i.e., aged 19 - 25 years) (Cancer Institute NSW, 2015). Although melanoma is the most lethal form of skin cancer, if detected early, the chance of someone living at least 10 years after the cancer is diagnosed is 99%, for individuals 60 years old or younger (Gimotty et al., 2007). Therefore, awareness of melanoma detection is important in preventing premature deaths from skin cancer.

2D Image Analysis

Conventional skin cancer detection methods such as physical inspection by a dermatologist is time-consuming with low accuracy (60-80%) (Saba et al., 2019). With constant improvement in computer technology, new skin cancer detection methods that are more cost-effective and accurate apply computer vision techniques such as image analysis. One of the crucial steps in these new methods is feature extraction. This process extracts important features that correctly emphasizes the key areas of skin lesions while lowering computational cost and increasing accuracy (Saba et al., 2019). In the case of 2-D image analysis, the features commonly used are color, texture and shape.

Similarly, other research has also shown that diagnosis of skin cancer can be automated using certain characteristics of various categories of skin cancer such as

color and feature information (Hoshyar et al., 2011). Main prognostic and diagnostic parameters of skin cancer melanoma are the shape, color, non-uniform pigmentation and irregular pattern on the lesion boundary (Hoshyar et al., 2011). Additionally, image or image sequence object recognition is also a crucial task (Singh et al., 2018). Although previous challenges such as multi-pose model recognition and multi-modal models have been dealt with by using 2-D objects, few studies in skin cancer classification have focused on proper feature extraction methods (Singh et al., 2018). We aim to also utilise feature extraction methods within the pre-processing stage of classification, specifically dull razor function, soft attention mapping and blurring images. The dull razor is a function that removes dark hair from input images allowing the CNNs to focus on the lesion, it is widely used in skin lesion analysis (Majumder & Ullah, 2019). Another common pre-processing function is soft attention mapping, where it improves classification performance of the model by guiding the model to focus on relevant areas and salient features of the input image (Figure 1) (Datta, et al., 2021). Lastly we also used blurring, where it is used to reduce noise from the input image and also increase accuracy in classification (Masood & Al-Jumaily, 2013).

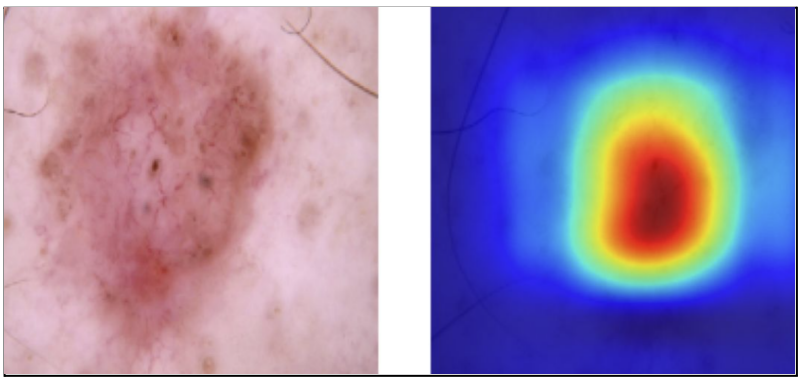


Figure 1. Example of soft attention mapping on a lesion

Currently, there are few machine learning models which can accurately detect skin cancers. Related literature who have used HAM10000 for skin cancer classification have obtained validation accuracies of 83% (Chaturvedi et al., 2020) and 91% (Khan et al., 2021). We are expecting to achieve a similar accuracy, however due to the novel approach of classifying against three classes, rather than the seven classes recorded in related literature, the accuracy of our project may be varied. Additionally, in literature most models are based on pre-trained CNNs, making model training

possible with a limited number of training data. Even so, the classification accuracy of these model types are still limited by the shortage of malignant lesion images. Two of the methods that could overcome this issue are 1) increase the malignant tumour images via image augmentation method 2) by using ensemble-based CNN architecture (Qureshi & Roos, 2021). We opted to forgo the second option due to memory size restrictions in model training.

To achieve robust and reliable results with limited resources and time, using different transfer learning models such as ResNet, VGG-16 and DenseNet have been recommended by previous researchers (Shawon et al., 2021). Researchers also proposed these five stages to the CNN models: 1) image acquisition 2) image preprocessing with hair removal using dull razor method before using median filter to smooth out the images, 3) image segmentation 4) feature extraction and 5) classification (Shawon et al., 2021). Although recent research results have again proven the superiority of CNN over other compared machine learning methods, this method is still not perfect **(**Subramanian et al., 2021). Image changes such as rotations or zooms can have a bad effect on the robustness of the performance of CNN in classifying skin lesions. However, these effects can be reduced via different methods but not fully eliminated. Further research is required to sustain CNN robustness and reliability prior to full adoption in the real medical industry (Maron et al., 2021).

Conclusion

In the age of artificial intelligence, scientists have proven cases of CNNs being better at detecting skin cancer lesions than most dermatologists (Haenssle et al., 2018). Overall, there are many successful pre-trained CNN techniques and attempts at analysing skin cancer image datasets. However, few studies focus on utilising various pre-processing methods such as dull razor, soft attention mapping and blurring images. It is therefore important to utilise the techniques of different pre-trained CNN models used in other skin lesion classification tasks and also incorporate the multi-modal aspect of feature pre-processing and feature extraction. Additionally it is also important to utilise our novel approach of using a three class classification, rather than the default seven class classification. Although CNNs have shown to be successful in detecting skin lesions accurately, medical history and

thorough examinations are also required to capture the full context of the patient. CNNs are useful to use as an aid to dermatologists in reducing workload.

3. Research/Project Problems

3.1 Research/Project Aims & Objectives

The project aims to build a classification to detect skin cancer using visual images of skin lesions. The skin lesion will be classified into one of the three classes ('mel', 'bcc', and 'others') and a probability score of the likelihood of belonging to the classified class compared to the other classes will be shown.

There are various factors that influence the risks of getting skin cancer, seen below (Cancer Institute NSW, 2015).

- Genetic and family background
- Age and sex
- Number of moles
- Geographical region
- Hair and skin colour

Due to project scope, the above features are excluded and considers only the patterns in the skin lesions. While some lesions are cancerous, others are harmless and benign, and do not require any treatment. The focus is to differentiate the skin lesions that are cancerous from ones that are harmless. We plan to use the 2D images to identify patterns for classification.

3.2 Research/Project Questions

With the objectives in mind, the project aims to answer whether we can use machine learning techniques to differentiate the patterns in melanoma to other skin lesions.

The ABCDE rule of melanoma diagnosis is a popular diagnosis technique used by specialists and medical professionals to diagnose skin cancer. (Centers for Disease Control and Prevention, 2020)

- Asymmetry: Melanomas are asymmetrical and one half of the melanoma won't resemble the other.

- Border: The edges are uneven, ragged, notched, or blurred while the moles have smooth edges.
- Color: Melanomas may have multiple colors. While moles have a uniform shade of brown color, melanomas may have different shades of colors including brown, black, red, white or blue.
- Diameter: Melanomas are usually larger in diameter than a mole.
- Evolving: Melanomas tend to change in terms of colour, size, and shape. This will not be a factor for 2D images as it is taken at a singular point in time.

Some of the above patterns usually exist in melanoma. There are some popular CNN models with different architecture that have been implemented by researchers and have been proven to be successful (Gouda & Amudha, 2020; Hosny et al., 2019). The project will implement similar techniques on the novel 3-class approach and will aim to enhance the already proven and established techniques, hopefully to the point where it will surpass traditional diagnosis techniques.

3.3 Research/Project Scope

Due to a level of uncertainty over the exact type of images that the client will utilize the model on, the scope of the project is focused more on enabling the client the ability to train the appropriate models with features for preprocessing, weights, and class options, as well as providing example models which work well on public skin lesion datasets.

At a minimum, the project scope has the following features and outcomes:

- The ability to accurately predict the class of the skin lesion based on 2D images.
- Output the prediction in terms of probability, inferring model confidence in the prediction made.

As a minimum, the final deliverable of the project would have the following features and outcomes:

- Ability to accurately predict the class of the skin lesion based on 2D images.
- Output the confidence of our prediction in terms of probability. Higher probability would mean that our model has more confidence about the prediction it has made about the input lesion belonging to the class.

4. METHODOLOGIES

4.1 Methods

CNNs are a proven technique for analysing visual images, it has been widely used in medical image analysis and has the ability to learn features from the images without needing to be built from scratch (Tajbakhsh et al., 2016). Consequently, it requires a large volume of annotated data for training. Given that publicly available data related to skin cancer is limited, another technique we can use is "transfer learning". Transfer learning is a process in which we pass the input data through the layers of a pretrained model and add additional layers as an extension to the model (Marcelino, 2018). Either all the layers in the model are trained or only the additional layers are trained.

4.1.1 Approach for 2D Images

The input dataset is preprocessed using various techniques such as dull razor function, blurring and others (explained later in this paper), before they are fed to the subsequent layers (Figure 2). The preprocessed image can then be segmented by segregating lesions from the background skin using soft attention mapping. The images are then passed to the classification layer in which the classification models are trained by fine tuning hyperparameters. The final model is chosen by evaluating performance metrics and a confusion matrix generated using a test dataset.

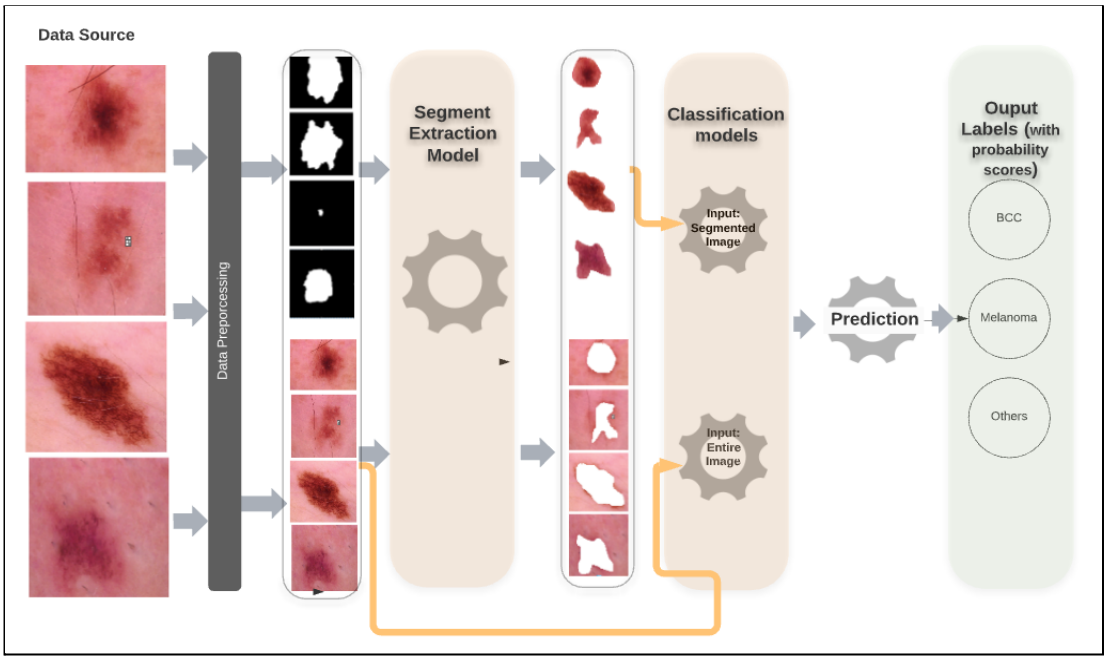


Figure 2. Overview of model training process.

4.1.2 Preprocessing

The input data was originally planned to follow a process of image size standardization, segmentation, and brightness normalization. While segmentation was implemented through usage of soft attention, it was generally found to produce worse results, although it is still possible to implement in the current model framework at a significant computational cost. That said, the prior process of size standardization and normalization is still applied within the process.

Image size standardization and testing of various resolutions was performed as planned, with most of the models settling for (512,512) resolution images. An addition to the preprocessing is the dull razor algorithm, designed to remove hair from the images such that hair would not become a factor of the image. The process involves the identification of dark hair based on segmentation on a grayscale image, creating a mask, before removing the relevant pixels using the mask, replacing them with pixels interpolated from its surroundings. This is usually followed by a smoothing blur. While the removal of the hair could potentially assist the model creation, the process also had the tendency to remove features that may be useful for the model. In Figure 3 below, while the hair visibility has been decreased substantially, so have the dots within the middle of the lesion, being reduced to grayish blurs.

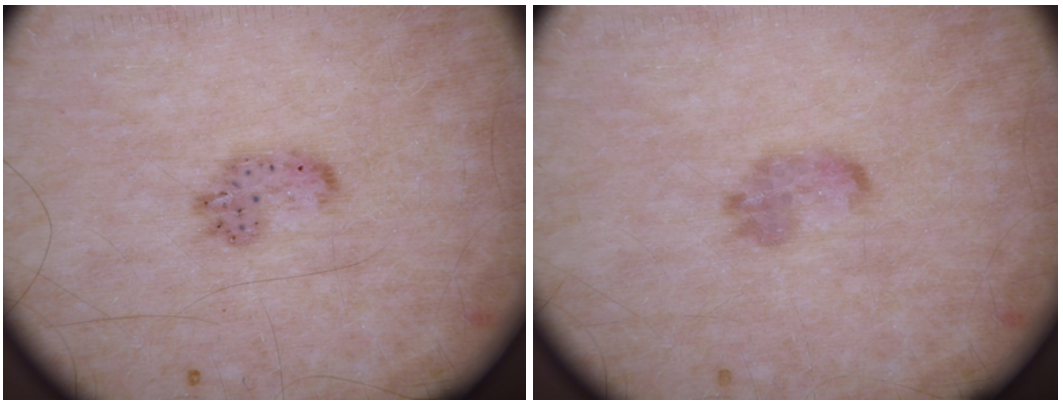


Figure 3. Comparison of raw and preprocessed images.

Soft attention utilized ResNet50, using the pretrained model to output a gradient; a heatmap for the location of the mole. This would ideally enable the model to prioritize certain parts of the image when training, easing the training process. Using the output to obtain a segmentation mask, this was then applied to the base image.

Dull razor as well as soft attention processing were initially implemented into the framework. This meant that each image was converted everytime it was loaded into the model training framework, creating significantly longer training times. This was remedied through creation of separate preprocessed datasets which would be fed into the training model, at the cost of available memory.

4.1.3 Model Training

1. **Class Imbalance:** The input dataset, HAM10000, is highly biased towards non-cancerous images with a ratio of 1:8 for cancerous to non-cancerous images. As a first step, the class imbalance in the dataset is fixed by upsampling the images in the class with fewer samples. Upsampling is achieved using augmentation techniques like rotation and flipping. As part of fixing class imbalance, all the classes are upsampled to the same size.

2. **Segment extraction model:** Models were trained to extract the skin lesion and the surrounding skin. The input data are the original images and the output is the segmentation masks generated in the preprocessing step. Using the output of the trained model, segmentation was done on new unseen data (i.e., test data) and the segmented parts (i.e., lesion and the surrounding skin) were passed as inputs to the classification models. As noted earlier, the segmentation did not yield better results and as it was computationally costly to implement, the approach was only used in the models that were trained initially and was later discarded.
3. **Classification models:** Classification models were trained within two contexts:

   a. **Training using the entire input image**: This approach used raw images as input data. In order to improve training performance, the images were standardised and normalised.

   b. **Training using the segmented data**: The inputs from the segment extraction model were passed to separate classification models. This approach was then discarded considering computational costs and no significant improvement in model performance.

A framework has been built with the following capabilities to automate the training process:

- **Configurable hyperparameters:** All the hyperparameters are made configurable using YAML based configuration files (Figure 4).

```yaml
ham_10k_resnet50_512v2:
  model_name: "HAM_10k_ResNet50_512v2"
  input_params:
    input_size: (512,512)
  model_params:
    batch_size: 32
    arch: tf.keras.applications.ResNet50
    freeze_pretrained: True
    steps_per_epoch: 100
    metrics: ['accuracy']
    loss:
      func: "sparse_categorical_crossentropy"
    optimizer:
      func: "Adam"
      params: "{'learning_rate':1e-4}"
    class_weight_mu: 1
  attention_config:
    resize:
      resizeW: 224
      resizeH: 224
    dull_razor:
      enabled: True
      razorblur: "M"
      mediankernel_razorblur: 3
      filterstructure: 5
      lowerbound: 5
      inpaintmat: 3
    blur:
      enabled: True
      normalblur: "M"
      mediankernel_blur: 5
      blurnum: 5
    soft_attention:
      alpha: 0.7
      beta: 0.3
      gamma: 0.0
```

Figure 4. A sample configuration file used for training

- **Plugin any pretrained model or use custom architecture for training**: In order to leverage on the architectures proven to work on image datasets, the framework has a functionality to plugin any popular architectures like ResNet, DenseNet and others. The pretrained models can be used and extended with additional layers.

- **Ability to freeze or train layers from pretrained models:** The default layers from the pretrained models can be frozen if only the additional layers need to be trained. This is achieved by setting the “freeze_pretrained” parameter in the configuration file to be “True”.

- **Save and load models:** The models that are trained will be saved after the first epoch to a disk in the form of Keras supported h5 file. Models will also

be saved after each epoch if the validation loss of the current epoch is better than the validation loss from any previous epochs. As part of saving a model, the model architecture, weight parameters, validation losses for each epoch and the state of the optimizer are all saved. The saved model can be reloaded for resuming the training or for predicting class labels given image inputs. This ensures the state of the model is completely saved and can be reused.

- **Prevents model overfitting by saving only the best epoch based on validation losses:** As only the models having the least validation losses are stored to the disk, it ensures overtraining doesn't affect the model performance.

- **Transfer learn using trained models on different datasets:** As the best model is saved as part of the training process, it can be loaded and used on a different dataset of skin cancer images for transfer learning or for validating the model against an external dataset that's not used for training. This will help us understand how well our model will work on a real work scenario or on a dataset that the model has never "seen".

- **Ability to add weights to particular classes:** It is important to have good recall for cancerous images as misclassifying a cancerous image as non-cancerous is costlier than misclassifying a non-cancerous image. To penalise the misclassification of cancerous images, adjustable weight parameters have been added to the framework that will set how misclassification of each label needs to be penalised.

- **Label predictions with probability scores:** Labels are predicted with probability score for each label. Using the probability score, the confidence on the prediction can be determined. This helps in building the risk profiler as the scores tell how certain the model is about the prediction.

- **Report performance metrics and confusion matrix:** Measuring the performance of the model at certain intervals reveal when the model training needs to be stopped. After each epoch, the accuracy on the validation dataset

is returned. In addition, the framework also reports other performance metrics within the confusion matrix at a set interval.

- **Variety of preprocessing methods available:** The framework has the capability to take any custom preprocessing function as input in addition to keras supported functions. The custom preprocessing function can be passed a preprocessing parameter and the images are preprocessed for each step before they are fed to the model training.

- **Reduces learning rate on plateau:** To fine tune the performance of the model and to achieve higher accuracy, the framework automatically reduces the learning rate when it sees a plateau in the validation loss surface.

4.1.4 Performance Metrics

Sensitivity and specificity are important metrics in a medical diagnosis. Sensitivity is the proportion of those who have the disease being correctly identified as having it (i.e., true positive), while specificity is the proportion of those who do not have the disease being incorrectly identified as having it (i.e., false positive). In addition to that, we will also be reporting other metrics of the models such as average accuracy, precision, recall and f1 score which will be made suitable for multi-class classification.

4.2 Data Collection

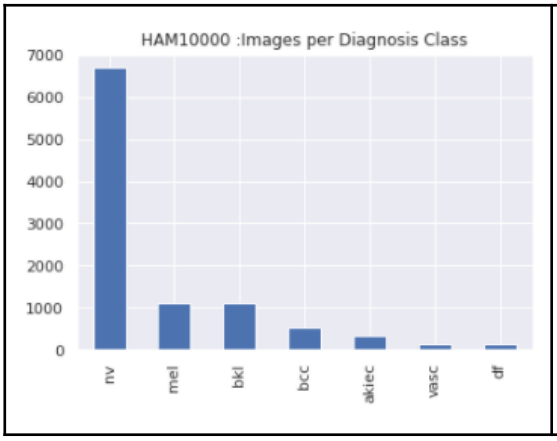
The main data source for the 2D image model is HAM10000, which originally was collected from different populations set in Austria & Australia over the period of 20 years (Tschandl et al., 2018). It was stored in different formats like powerpoint files, diapositives, etc and various extraction techniques used to consolidate the data finally in jpg format .It was extracted from Kaggle for the purpose of this project. It includes a representative collection of all important diagnostic categories in the area of pigmented lesions.

The other two sources of data are ISIC-2019 (Codella et al., 2017; Combalia et al., 2019; Taschandl et al., 2018) and PAD-UFES-20 (Pacheco et al., 2020). ISIC-2019 consists of below two datasets in addition to HAM10000:

- **BCN20000**: It consists of 19424 dermoscopic images of skin lesions captured from 2010 to 2016 in the Hospital clinic of Barcelona Spain
- **MSK**: Dataset includes over 12500 images initially used in ISIC 2018 challenge.

4.3 Data Analysis

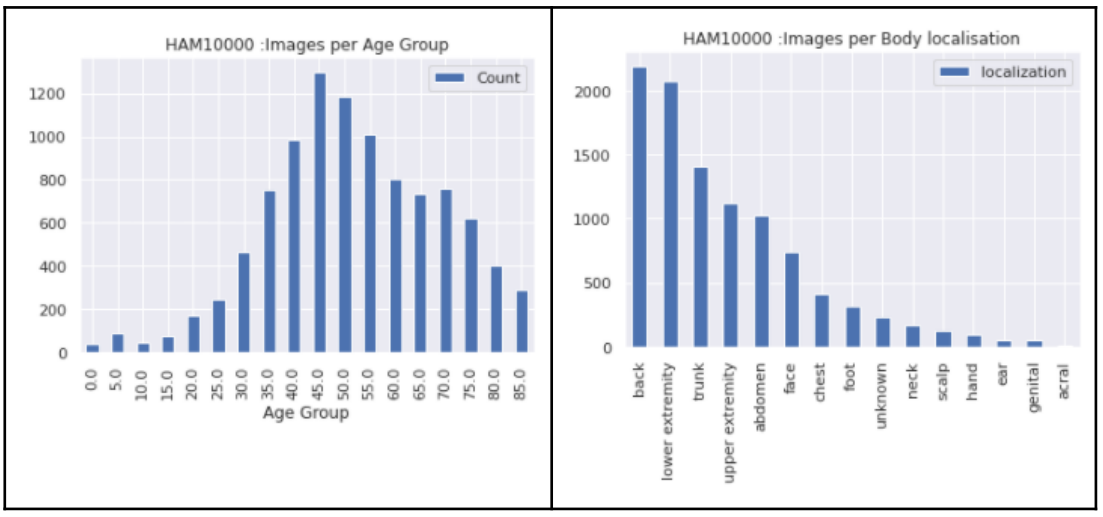
HAM10000 Data is extracted from Kaggle and read through the Jupyter Notebook. Actual images are stored on the hard disk of GCP/AWS as jpeg format with an image ID. This ID is linked to the metadata csv file which stores other features information of Lesion image discussed below. Total 10,015 images are distributed in below 7 classes as shown below (Figure 5).



Skin Lesion type	Acronym	Count
Melanocytic nevi	nv	6,705
Melanoma	mel	1,113
Benign keratosis	bkl	1,099
Basal cell	bcc	514
Actinic Keratoses (Solar Keratoses) and Intraepithelial Carcinoma (Bowen's disease)	akiec	327
Vascular skin lesions	vasc	142
Dermatofibroma	df	115

Figure 5. Counts of images per class.

Exploratory analysis of the features using the metadata csv file was conducted and below charts in Figure 6 were produced using Python matplotlib & Seaborn libraries.



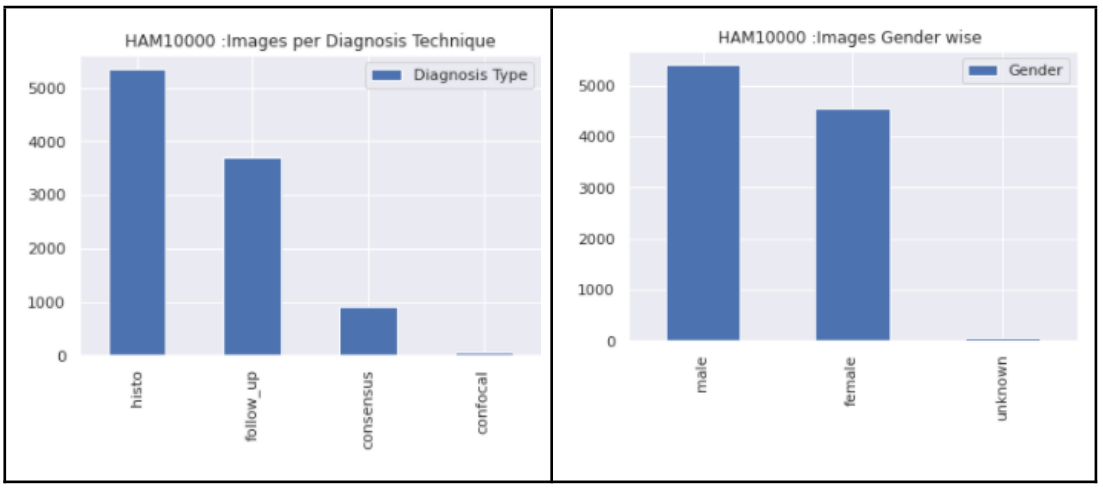


Figure 6. Exploratory data analysis on HAM10000.

Below key points are noted based on the above bar charts generated from HAM10000 dataset:

- There is a drastic increase in lesion occurrences whether cancerous or benign from the age 35. Most of the people impacted are between the age group of 35 to 60 yrs
- Half of the total occurrences of lesions are identified through histopathology which is a costly process and needs specialist advice (Shi et al., 2020).
- Back, lower / upper extremities and trunk are the major locations where most of the lesions are found.
- Males were found to have a higher number of lesions as compared to females in HAM10000 dataset

On similar lines, ISIC2019 and PAD-UFES-20 images were also explored and class distribution shown below in Figure 7.

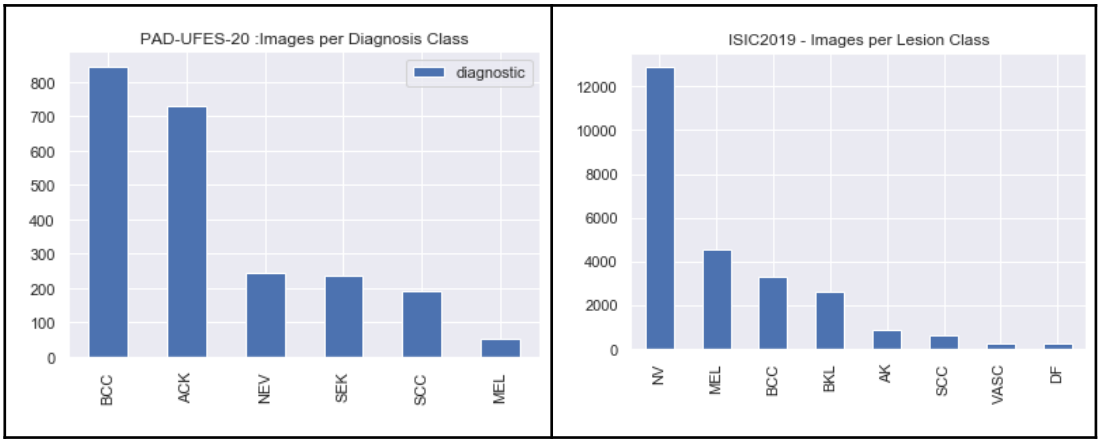


Figure 7. Exploratory data analysis on ISIC2019 and PAD-UFES-20.

In general all 3 datasets suffer from class imbalance problems which were addressed in pre processing before model training.

5. RESOURCES

5.1 Hardware & Software

Programming Language	Python 3.8
Python Libraries	numpy, pandas, matplotlib, openCV, sklearn, yaml, dataclasses, dacite, math, json, itertools, os, pathlib, cv2, seaborn, skimage, random
Deep Learning Framework	Tensorflow 2.4
Deep Learning APIs	Keras, Tensorflow Low Level APIs
Training Platform	Google Colab with GPU/TPU, Google Cloud Platform Deep Learning VM, and Amazon Web Services EC2 instances
Data storage	Google Drive, Google cloud storage
Source Code Management	Github
Team collaboration	Slack, WhatsApp, Zoom

5.2 Materials

Hardware: Models analysing images require more parameters, therefore training them in systems with GPUs with sufficient memory helped us do more iterations which in turn helped us improve the overall performance of the models. The Google Cloud Platform and Amazon Web Services have been utilized for the purpose of meeting the hardware requirements for model training.

6. MILESTONES/SCHEDULE

Milestone	Tasks	Reporting	Date
Week-1	Selecting a team and top three topic preferences	Team and Topic Selection	1-03-2021
Week-2	First Client and Tutor meeting on March 19th Researching papers for CNN and risk factor models on classifying skin cancer Research relevant pretrained models for usage in the 2d image analysis model	ASANA collaboration tool	8-03-2021
Week-3	Tutor meeting on March 17th Client meeting on March 18th Started creating CNN model, framework, pre-processing and feature extraction. Started coding work.	Individual Reports, ASANA collaboration tool	15-03-2021
Week-4	Tutor meeting on March 24th Client meeting on March 23rd Research 3D image datasets and 3D image analysis procedures	Individual Reports, ASANA collaboration tool, Github	22-03-2021

	Github creation, colab compilation Continue work on fine tuning pre-processing, dataframe, Alexnet, Resnet, Vggnet.		
Week-5	Proposal Report Due Continue work on fine tuning pre-processing, dataframe, Alexnet, Resnet, Vggnet.	Proposal Report, Individual Summary Report, Github	29-03-2021
Week-6	Tutor meeting on April 6th Client meeting on April 7th Research 3D image datasets and 3D image analysis procedures Continue work on fine tuning pre-processing, dataframe, Alexnet, Resnet, Vggnet Submit model to train on our selected cloud service	Individual Reports, ASANA collaboration tool	12-04-2021
Week-7	Tutor meeting on April 13th Client meeting on April 14th Finalize work regarding the 2d image model Get 3D image dataset from the client	Individual Reports, ASANA collaboration tool, Github	19-04-2021

	Get started with 3D image analysis and implementing it into the model Get required infrastructure from Tutor and Client		
Week-8	Tutor meeting on April 20th Client meeting on April 21st Get required infrastructure from Tutor and Client Train 3D images with our current models and find the optimal model. Get feedback from tutor and client about out model for us to make further improvement	Individual Reports, ASANA collaboration tool	26-04-2021
Week-9	Progress Report Due Tutor meeting on April 28th Client Meeting April 27th	Proposal Report, Individual Summary Report, Github Debug preprocessing and framework	3-05-2021
Week-10	Tutor meeting on May 4th Client meeting on May 5th Final Presentation and Final Report preparation	Individual Reports, ASANA collaboration tool Project demonstration to client	10-05-2021

Week-11	Tutor meeting on May 12th Client meeting on May 11th Final Presentation and Final Report preparation	Individual Reports, ASANA collaboration tool, Github	17-05-2021
Week-12	Final Presentation Tutor meeting on May 19th Client meeting on May 18th	Github	24-05-2021
Week-13	Final Report (thesis)	Individual Summary Report Send client deliverables	31-05-2021

7. RESULTS

7.1 ResNet50

Running the original layers of ResNet50 on the raw HAM10000 data resulted in a validation accuracy of 66%. In order to improve this accuracy, the training set was upsampled to correct imbalancing. The training set images were also normalised so each image falls within the same range, making gradients within the model more stable. In addition to the default ResNet layers, additional custom layers seen in table 1 below were also added and parameters were tuned to find an optimal accuracy as requested by the client. Furthermore table 2, shows the parameters used within training the model to find optimal accuracy and Figure 8, the total amount of parameters. The final model resulted in a validation accuracy of almost 90%.

Table 1. Final ResNet50 custom layers and parameters used to achieve 89% validation accuracy.

Layer name	**Parameters and values used**
Global average pooling	-
Flatten	-
Dense	512 nodes, relu activation and L2 regulariser at 0.001
Dropout	0.5
Dense	512 nodes, relu activation and L2 regulariser at 0.001
Dropout	0.5
Dense	3 nodes, softmax activation and L2 regulariser at 0.001

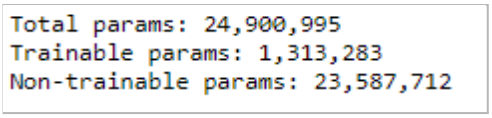


Figure 8. Total parameter count in the final ResNet50 model

Table 2. Parameters used to train the final ResNet50 for 89% validation accuracy.

Parameter name	**Values used**
Input image size	512x512

Batch size	32
Epochs	33
Loss function	sparse categorical cross entropy
Adam optimizer function	learning rate at 0.0001
Melanoma, BCC and Others class weights	Melanoma: 2.5, BCC: 2.5 and Others: 1

Due to the limited publicly available datasets, there are too few images of melanoma and BCC lesions (cancerous images), resulting in a highly imbalanced dataset. Although upsampling was performed on the training set, as seen in Figure 9 below, the model is still focusing on the majority group 'Others'. The recall scores for 'Others' class is over 90%, however the important cancerous classes 'bcc' and 'mel' are poor.

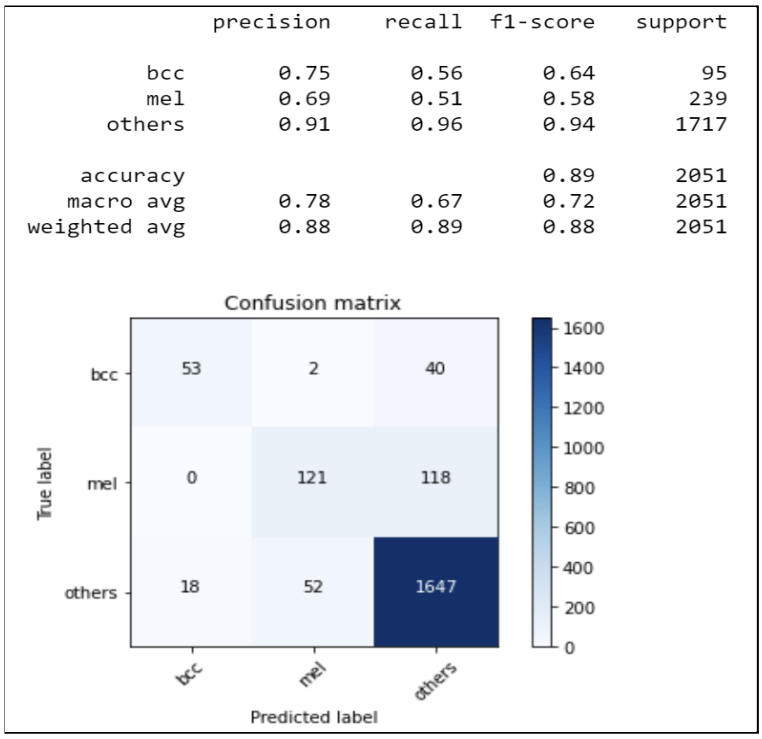


Figure 9. Test performance of ResNet50 trained on HAM10000 dataset resulting in 89% validation accuracy

Although the validation accuracy dropped by 7%, changing the weights of the classes, by putting more weight towards the cancerous classes (i.e., 'mel':3, 'bcc':3, 'others':0.5) resulted in improved recall scores of 80% for all classes, seen in the below confusion matrix (Figure 10).

	precision	recall	f1-score	support
bcc	0.50	0.82	0.62	95
mel	0.45	0.83	0.58	239
others	0.96	0.82	0.89	1717
accuracy			0.82	2051
macro avg	0.64	0.82	0.70	2051
weighted avg	0.88	0.82	0.84	2051

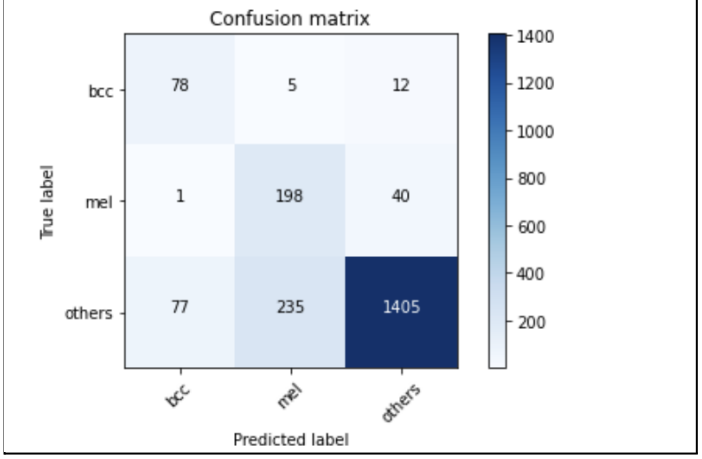


Figure 10. Test performance of ResNet50 trained on HAM10000 dataset resulting in 82% validation accuracy, however greater recall scores than the previous model.

7.1.1 Testing custom weighted ResNet on new data - ISIC-2019

After training ResNet50 on HAM10000, the weights of the model were saved and used as custom weights to train and test a new dataset called ISIC-2019 (Codella et al., 2017; Combalia et al., 2019; Taschandl et al., 2018). The results are shown in the below confusion matrix (Figure 11).

	precision	recall	f1-score	support
bcc	0.63	0.80	0.70	652
mel	0.59	0.69	0.64	947
others	0.89	0.80	0.84	3509
accuracy			0.78	5108
macro avg	0.70	0.76	0.73	5108
weighted avg	0.80	0.78	0.79	5108

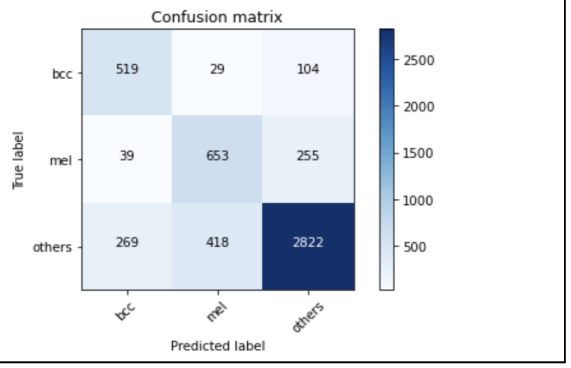


Figure 11. Test performance of custom weighted ResNet50 trained on ISIC-2019 dataset resulting in 78% validation accuracy.

As requested by the client we also created a function which inputs one external image using the custom weighted model and outputs a prediction of what class the image belongs to including the probability of the image belonging to the class. The image was not trained by the model prior. An example below (Figure 12) shows that the lesion melanoma was correctly identified by the custom model and the output showed that the probability of the image being melanoma is 0.96. It also showed the probability of the lesion being the other classes which were close to 0.

In [15]:
```
#image file directory goes here
image = r'/home/ubuntu/data/isic-2019/isic_train_by_class/mel/ISIC_0000002.jpg'

#function which outputs probability scores of all classes and highest probability class
label, label_prob, all_labels_prob = model_trainer.predict(image)

print("Probability scores range between 0 and 1. Higher scores indicate more probability of lesion belonging to the class.""

print("\033[1m""Lesion is most likely", label, "class with a probaility of", label_prob,"\033[0m""\n")
print("All class proabilities for this lesion:",
      f"bcc - {all_labels_prob[0][0]}, mel - {all_labels_prob[0][1]}, others - {all_labels_prob[0][2]}")
```

```
Probability scores range between 0 and 1. Higher scores indicate more probability of lesion belonging to the class.

Lesion is most likely mel class with a probaility of 0.96491456

All class proabilities for this lesion: bcc - 2.885732646973338e-05, mel - 0.9649145603179932, others - 0.03505659103393555
```

Figure 12. The output of only testing an external image into the custom weighted ResNet model.

7.1.2 Testing custom weighted ResNet on new data - PAD-UFES-20

Using ResNet50, the weights of the model were saved and used as custom weights to train and test a new dataset PAD-UFES-20 (Pacheco et al., 2020). These images are smartphone images and it resulted in a 0% recall score for melanoma images. Overall, the model resulted in 74% accuracy, seen in Figure 13.

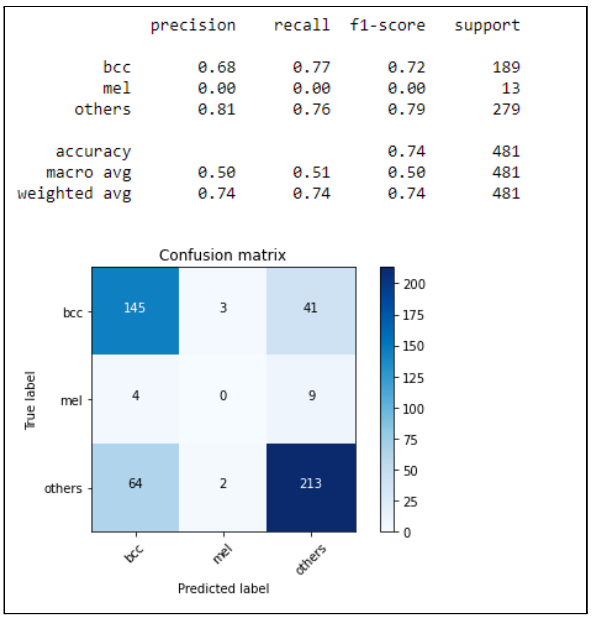


Figure 13. Test performance of custom weighted ResNet50 trained on PAD-UFES-20 dataset resulting in 74% validation accuracy

7.1.3 Experimentation of ResNet on preprocessed data

We applied the dull razor algorithm and blur on the datasets and created alternate preprocessed versions of the augmented training data (Figures 14 to 16). This was also further performed on soft attention mapping, a method that was assessed to be detrimental. We also preprocessed the relevant test data using the same preprocessing settings of the training data to prevent information leakage. Additionally, we utilised Keras's inbuilt preprocessing to center and standardize images, as using custom preprocessing framework would require implementation of it manually. Experimentation regarding class weights and preprocessing variables was performed.

Trialling various preprocessing variables was initially done on a reduced training dataset of 2000 images per class due to the amount of time it would take to preprocess, while still allowing for the production of quantitative results rather than simply observing converted images and assessing the quality of the preprocessing. The best performing variables were then used to convert the full dataset. While the

majority of this was performed using the resnet model with custom layers, this was initially performed on densenet. It was from this that attention mapping was determined to be detrimental to the model, producing accuracies and recalls significantly worse than the preprocessed and dull razored versions.

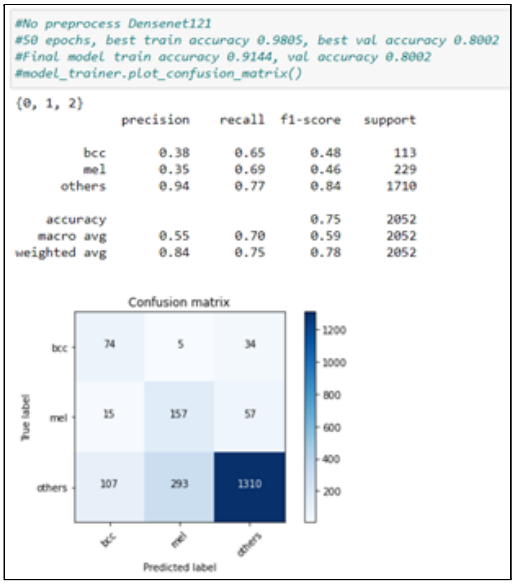


Figure 14. Initial analysis and assessment of performing no preprocessing using training dataset size of 2,000 images per class and DenseNet.

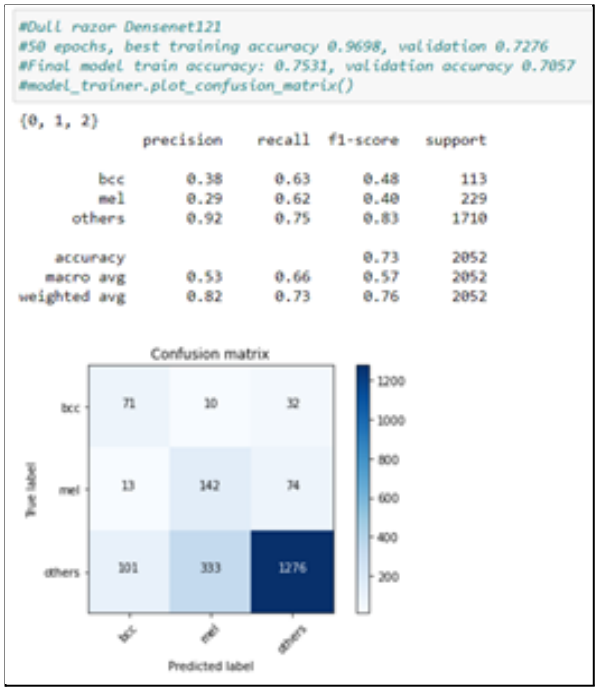


Figure 15. Initial analysis and assessment of dull razor algorithm using training dataset size of 2,000 images per class and DenseNet.

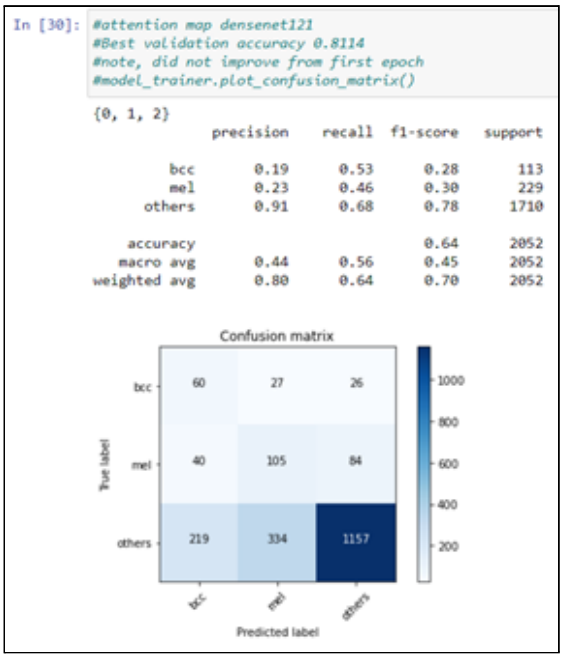


Figure 16. Initial analysis and assessment of soft attention mapping using training dataset size of 2,000 images per class and DenseNet.

The results on the full dataset (8500 per class) using dull razor pre-processing on the custom layers resnet model are shown below in the following confusion matrix (using weights of 'mel':2, 'bcc':2, 'others':1) (Figure 17) .

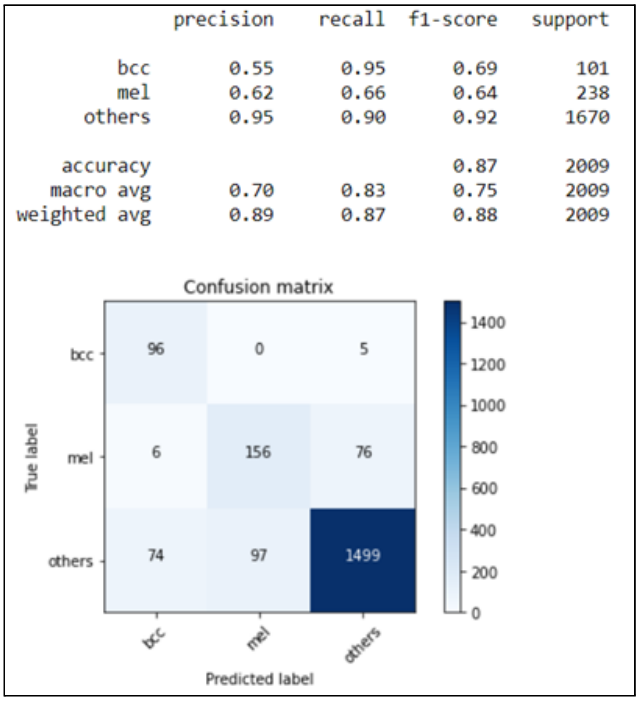


Figure 17. Test performance of ResNet50 trained on the preprocessed HAM10000 dataset using the custom weighted ResNet model.

In an attempt to correct for the low melanoma recall, weights were further modified to ‘mel’:3, ‘bcc’:2, ‘others’:1 (Figure 18).

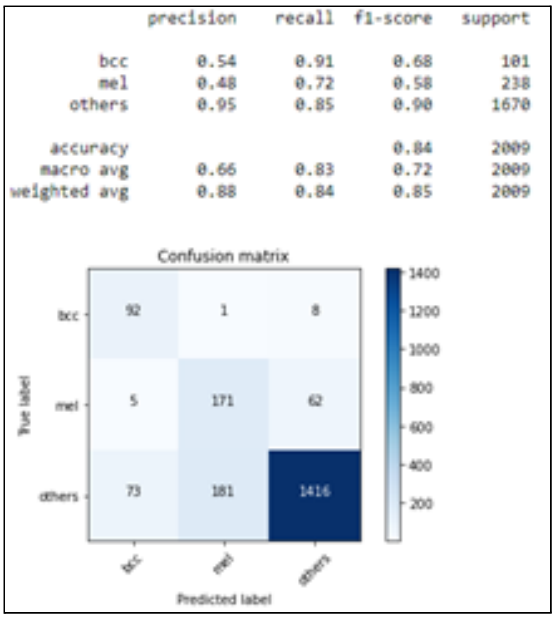


Figure 18. Test performance of ResNet50 trained on the preprocessed HAM10000 dataset using the custom weighted ResNet model with higher weighting for melanoma.

7.1.4 ResNet performance on 7 classes

As per client request, a model was created for HAM10000 using all available seven classes. Below are the results (Figure 19).

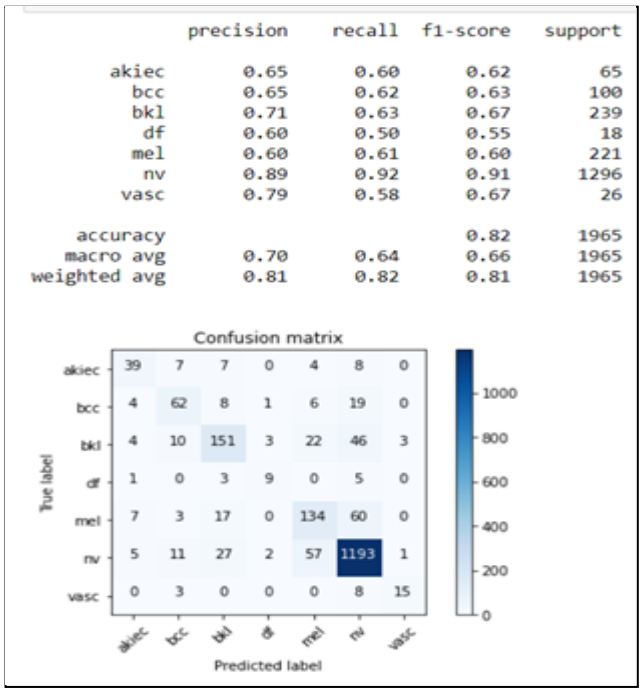
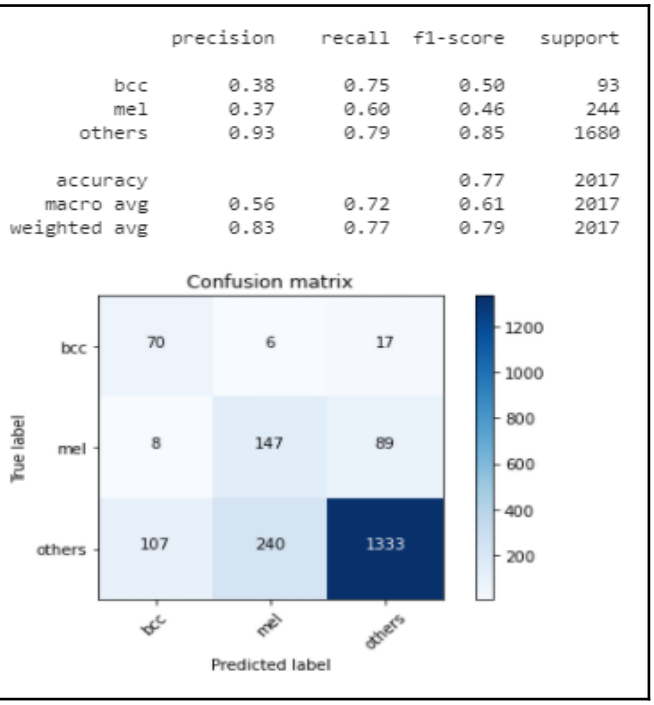


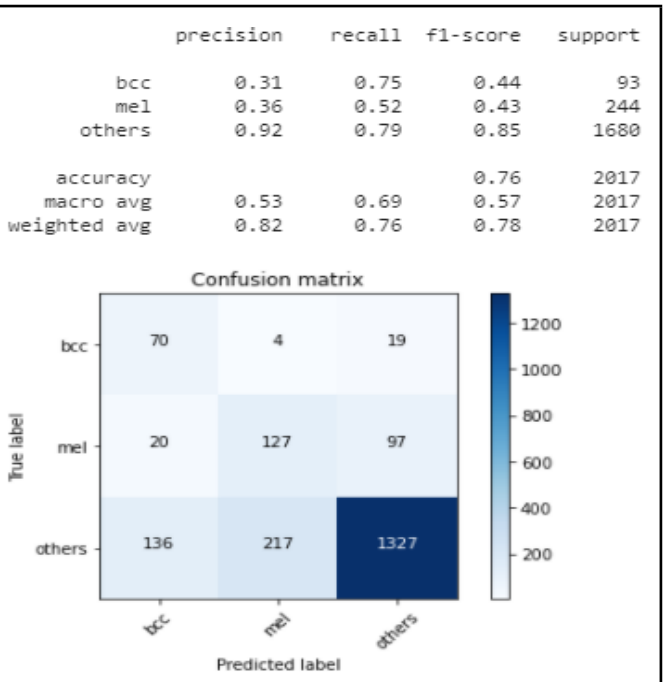
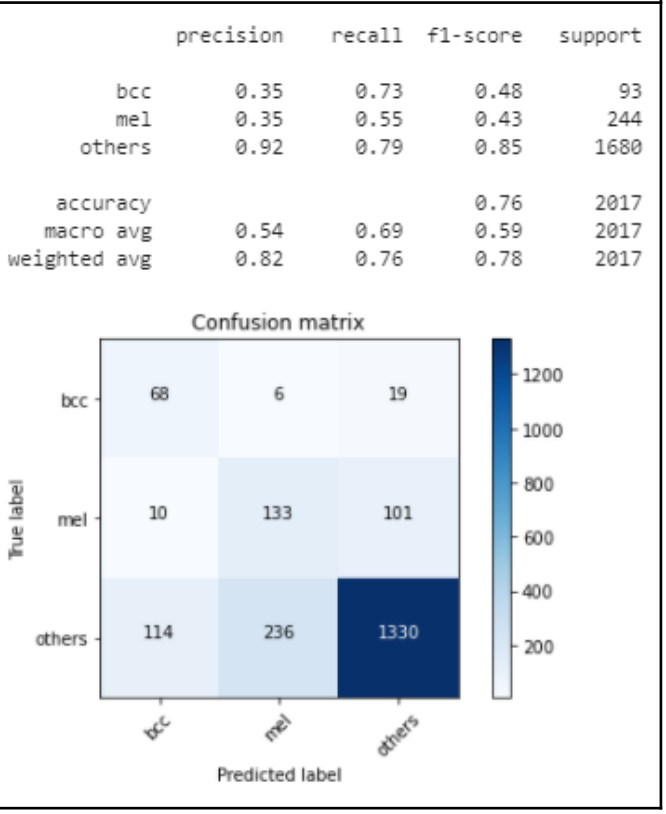
Figure 19. Test performance of ResNet50 trained on the HAM10000 dataset using the custom ResNet model on all classes in the base dataset.

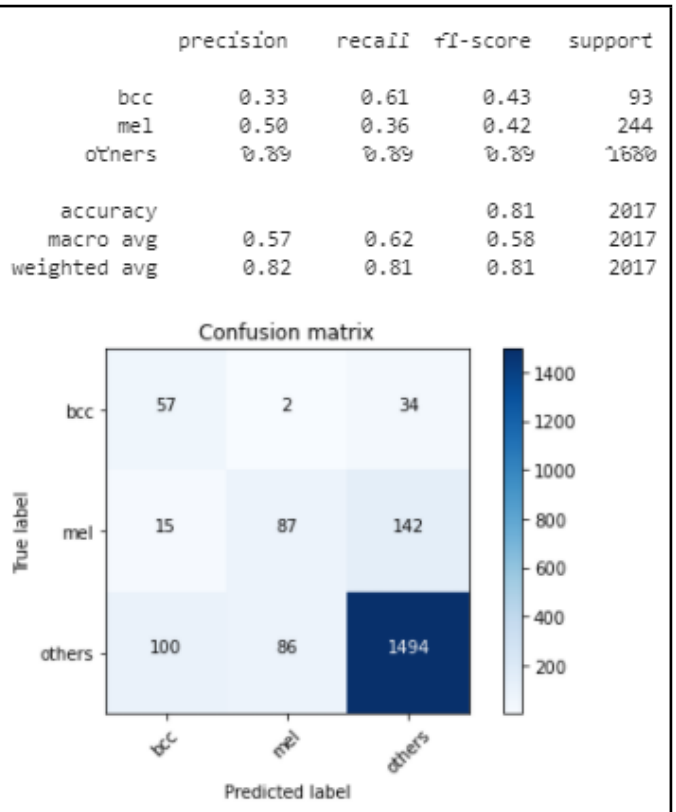
7.2 Experimentation with Xception and Inceptionv3

HAM10000 training dataset was used to train model with Xception & InceptionV3 architectures with different hyperparameters but validation accuracy did not go more than 81% .Below are the results from confusion matrix attached along with validation accuracy, architecture, epoch and other parameters (Table 3).

Table 3. Experimentation with Xception & Inceptionv3 CNN architectures.

Inception-V3 val_accuracy - 0.77 Epoch -35 Batch size -32 Learning rate - 1e3 Image Resolution (224*224)	

Xception val_accuracy - 0.76 Epochs - 30 Batch Size -16 Learning rate - 1e4 Image Resolution (224*224)	
InceptionV3 val_accuracy: 0.76 Epoch 30 Batch Size-16 Learning rate - 1e3 Image Resolution (224*224)	

Xception val_accuracy: 0.81 epoch 30 Batch Size -32 Learning rate - 1e4 Image Resolution (448*448)	

8. Discussion

The aim of the project was to build a classification model to accurately detect images of skin lesions, differentiating the lesions from cancerous types to non-cancerous. As requested by the client to focus on overall accuracy of the model, we achieved the requested near 90% accuracy of the model that was trained and tested on HAM10000, at the cost of recall scores. However, putting more weight towards the cancerous classes 'mel' and 'bcc' allowed the model to improve recall scores as follows. For bcc skin cancer images 82% of actual bcc images were correctly identified as being bcc, 83% of actual melanoma images were correctly identified as being melanoma, and 82% of other skin lesions were correctly identified as being other skin lesions. Other literatures who also used HAM10000 to classify skin lesion images obtained validation accuracies of 83% (Chaturvedi et al., 2020) and 91% (Khan et al., 2021), however it is important to note that these authors utilised their CNNs to classify on the default seven classes rather than the current study's novel approach of three class classification. Therefore comparisons of these results may not be appropriate.

Furthermore, we also achieved the next request by the client which was to detect an external image at a time and classify the class it belongs to, with also including the probability of the lesion being the predicted class and the probability of the lesion belonging to the other two classes. Additionally, our method improves recall with images taken via dermatoscopes only. We tested our trained model on smartphone images (i.e., PAD-UFES-20) and produced poor results. With limited resources, further research and time is required to refine our preprocessing steps in order to broaden the scopes of our dataset.

Similarly, regarding dull razor, despite issues with testing preprocessing variables, reasonable parameters were found that produced interesting results, with a high of 95% of bcc images being correctly identified. Weighting in an attempt to balance the classes produced recalls of 91% on bcc, 72% on melanoma, and 85% on other classes. In relation to the main model, the noticeable prioritization of bcc over melanoma despite having equal class weights to melanoma was interesting, especially when considering the precision of bcc was generally better as well. An implication to this is that preprocessing of the data resulted in a model that could better predict bcc at the cost of other classes, and may be of use if a model stack were to be implemented. Attention mapping, while discarded early on, may still have potential, and perhaps the full extent of exploration regarding the process has not been explored. That said, another reason why attention mapping may not have produced comparable results was that it was drawing attention more to the center of the mole as opposed to the border and shape of the mole, or perhaps was simply not mapping around the mole correctly.

Compared to accuracy scores that were preferred by the client, we chose to focus on recall scores. It was understood by the authors of this paper that the project would be utilised as a simple early screening test by the client, and therefore it seemed reasonable that a prioritization of recall of cancerous classes would be more important. The cost of directing an individual to a skin cancer specialist who does not have skin cancer is far lower than the cost of failing to indicate to an individual that they do not have skin cancer when they do. As a result, rather than prioritizing correctly predicting all classes equally, focus was placed on recalls of cancerous classes.

Regarding misclassified images, for some it seems inevitable that they would be misclassified by both machine and human specialists. Below in figure 20 are two examples of melanoma images from the HAM10000 dataset. The left one is far less obvious; while there is some blurring regarding the border, colour is a fair bit more consistent and is fairly round. In contrast, the right one has obvious variation in colour, sporting purplish to black shades in addition to the brown as well as white discolorations, in addition to having unclear borders and being asymmetric. It should come as no surprise that the left image would be misclassified as a benign skin lesion belonging to the 'others' class, while the right image would be correctly classified as a melanoma.

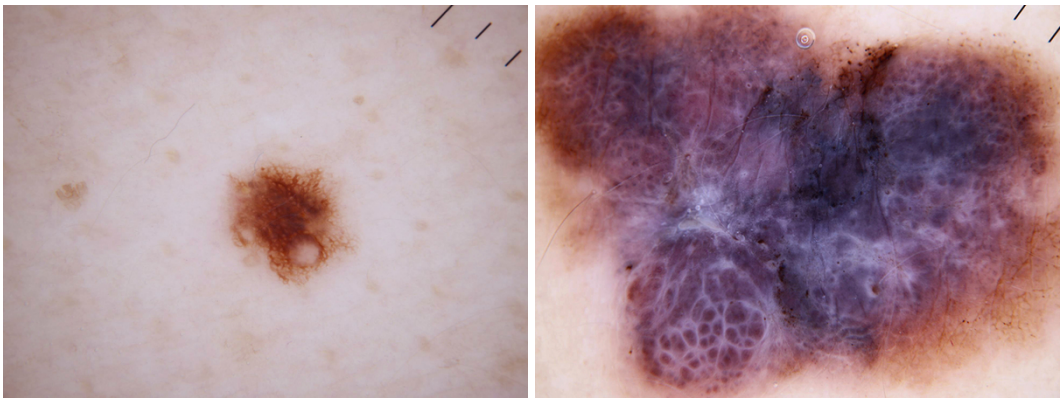


Figure 20. Example melanoma images from HAM10000.

The dataset that is used for training was of dimension 600 x 450. One of the key observations during the model training was that having images with good resolution improves the performance of the model as it can capture more details and information about the lesions. Given that the cameras in the mobile phones these days can capture images with high resolution, the prediction accuracy can be improved with better images. Moreover having multiple images of the same lesion will help improve the accuracy. Finally, the progression of the lesion is a critical indicator that's used for diagnosing skin cancer. Having images taken at different times (i.e., Sequential Digital Dermoscopy Imaging [SDDI]), during the progression of lesion may be useful to build a time sequence model which may yield better accuracy (Tschandl P., 2018).

9. Limitations and Future Works

Due to the limited public availability of datasets, there are too few images of melanoma and bcc lesions (cancerous images), resulting in a highly imbalanced dataset. Future studies should aim to find more cancerous images to upsample the two classes. As an example, early this year, Google has managed to create a state of the art trained model for skin cancer prediction achieving tremendous results (Financial Times 2021)). Google being a leader in technology has the vast resources to run the training and testing of over 1 million skin lesion images. This scale of these resources is beyond the reach of this project and other majority of companies until there are more free sources of smartphone dermatoscopic images available.

Similarly, another severe limitation was the lack of sample images from the client. It was uncertain as to what kind of images would be fed into the model. Simple smartphone images, or images taken via a dermatoscopic attachment to the smartphone. While this reduced efficiency, having to source from public datasets, the main issue that arose from this was the uncertainty regarding which images would be appropriate for the model.

Furthermore, due to time constraints and lack of sourcing 3D skin images, future studies should venture into 3D skin image analysis and the potential for a 3D classification model to be combined with 2D image analysis. Additionally, due to computational restrictions and limited memory and GPU allocation, we were unable to go forward with ensembling methods. Ensembling techniques have been widely used in machine learning models and this approaching method in skin classification appears promising for future work (Gessert et al., 2020).

In addition to the limited potential of ensembling methods, preprocessing was also severely hampered. Allowing preprocessing to run one by one, as the framework went through each image, resulted in training times that were up to ten times slower than without it. Putting the whole dataset through preprocessing once as an extra copy resulted in memory issues within the virtual machines available to us. As a result, assessing the preprocessing performance on images became difficult, as preprocessed datasets could only then be assessed qualitatively, or quantitatively on reduced datasets. In the future, better resources may allow for greater exploration and assessment of preprocessing variables, as well as production of models which may

work better for specific classes, perhaps further aiding an ensemble model if that were to occur.

Advanced methodologies can be explored in the future such as:

- **Dilated Convolutions**

  This method has generally improved performance as convolutions are applied to input with defined gaps. It increases the receptive view of the network exponentially without losing the resolution. In short , it allows a higher receptive field with the same computation and memory cost while retaining resolution (Yu. & Koltun, 2015). This will be significant for high resolution cancer images based on our experience.

- **Cyclical Learning Rate**

  We were limited in our approach of manually experimenting with different learning rates (which is a critical hyper -parameter for training CNN model) on 1 GPU machine. In this method learning rate varies cyclically between reasonable boundary values, and gets better classification accuracy in fewer iterations. (Smith ,2015).

- **Hyperparameter tuning via Grid search Pipeline**

  Hyperparameter optimization is a significant part of deep learning especially when there are a lot of parameters to be configured. Grid Search capability from scikit-learn library can be further explored to tune hyper parameters (e.g., learning rate, dropout rate, epochs and number of neurons) of Keras framework.

10. References

Ali, M. S., Miah, M. S., Haque, J., Rahman, M. M., & Islam, M. K. (2021). An enhanced technique of skin cancer classification using deep convolutional neural networks with transfer learning models. *Machine Learning with Applications, 5,* 100036–. https://doi.org/10.1016/j.mlwa.2021.100036

Argenziano, G., & Soyer, H. P. (2001). Dermoscopy of pigmented skin lesions: a valuable tool for early diagnosis of melanoma. *The Lancet Oncology*, *2*(7), 443–449. https://doi.org/10.1016/s1470-2045(00)00422-8

Australian Institute of Health and Welfare. (2015). *Leading cause of premature mortality in Australia fact sheet: Melanoma*. (Cat. No. PHE 202). Retrieved from: https://www.aihw.gov.au/getmedia/2be8b8ce-9ea8-447a-88f8-da60e1d0462f/phe202-melanoma.pdf.aspx

Cancer Institute NSW. (2015, December 8). *Skin cancer costs NSW upwards of $500 million* [Press release]. Retrieved from: https://www.cancer.nsw.gov.au/what-we-do/news/skin-cancer-costs-nsw-upwards-of-500-million.

Centers for Disease Control and Prevention. (2020). *What are the symptoms of skin cancer?* Retrieved from: https://www.cdc.gov/cancer/skin/basic_info/symptoms.htm

Chaturvedi, S. S., Gupta, K., & Prasad, P. S. (2020). Skin lesion analyser: An efficient seven-way multi-class skin cancer classification using MobileNet. In *Advanced Machine Learning Technologies and Applications* (pp. 165–176). Springer Singapore. https://doi.org/10.1007/978-981-15-3383-9_15

Chollet, F. (2017). Xception: Deep learning with depth wise separable convolutions. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1251-1258.

Codella, N. C. F. , Gutman, D., Celebi, E.M., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., & Halpern, A. (2017). Skin lesion analysis toward Melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the International Skin Imaging Collaboration (ISIC). *Arxiv e-prints*, arXiv:1710.05006.

Combalia, M., Codella, N.C.F., Rotemberg, V., Helba, B., Vilaplana, V., Reiter, O., Halpern, A.C., Puig, S., & Malvehy, J. (2019). BCN20000: Dermoscopic lesions in the wild. *Arxiv e-prints*, arXiv:1908.02288.

Datta, S.K., Shaikh, M.A., Srihari, S. N., & Gao, M. (2021). Soft-attention improves skin cancer classification performance. *MedRxiv, 2021.05.12.21257114.* doi: https://doi.org/10.1101/2021.05.12.21257114

Doran, C. M., Ling, R., Byrnes, J., Crane, M., Searles, A., Perez, D., & Shakeshaft, A. (2015). Estimating the economic costs of skin cancer in New South Wales, Australia. *BMC Public Health*, *15*, 952. https://doi.org/10.1186/s12889-015-2267-3

Financial Times. (2021). *Google launches AI health tool for skin conditions.* Retrieved from: https://www.ft.com/content/6d4cd446-2243-43f4-befd-565b4e880811

Gessert, N., Nielsen, M., Shaikh, M., Werner, R., & Schlaefer, A. (2020). Skin lesion classification using ensembles of multi-resolution EfficientNets with meta-data. *MethodsX*, *7*, 100864–100864. https://doi.org/10.1016/j.mex.2020.100864

Gimotty, P. A., Elder, D. E., Fraker, D. L., Botbyl, J., Sellers, K., Elenitsas, R., Ming, M. E., Schuchter, L., Spitz, F. R., Czerniecki, B. J., & Guerry, D. (2007). Identification of high-risk patients among those diagnosed with thin cutaneous melanomas. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*, *25*(9), 1129–1134. https://doi.org/10.1200/JCO.2006.08.1463

Gouda, A., & Amudha, J. (2020). Skin cancer classification using ResNet. *American Journal of Computer Science and Information Technology*, *8*(4), 52-58. https://doi.org/10.1109/ICCCA49541.2020.9250855

Haenssle, H. ., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., Kalloo, A., Hassen, A. B. H., Thomas, L., Enk, A., Uhlmann, L., Alt, C., Arenbergerova, M., Bakos, R., Baltzer, A., Bertlich, I., Blum, A., Bokor-Billmann, T., Bowling, J., … Buhl, T. (2018). Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology*, *29*(8), 1836–1842. https://doi.org/10.1093/annonc/mdy166

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, arXiv:1512.03385. https://arxiv.org/abs/1512.03385

Hoshyar, A. N., Al-Jumaily, A., & Sulaiman, R. (2011). Review on automatic early skin cancer detection. *International Conference on Computer Science and Service System (CSSS)*, INSPEC: 12226388, pp. 4036-4039. https://doi.org/10.1109/CSSS.2011.5974581

Hosny, K. M., Kassem, M. A., & Foaud, M. M. (2019). Classification of skin lesions using transfer learning and augmentation with Alex-net. *PLoS ONE*, *14*(5), e0217293. https://doi.org/10.1371/journal.pone.0217293

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, arXiv:1608.06993. https://arxiv.org/abs/1608.06993

Khan, M. A., Sharif, M., Akram, T., Damaševičius, R., & Maskeliūnas, R. (2021). Skin lesion segmentation and multiclass classification using deep learning features and improved moth flame optimization. *Diagnostics (Basel)*, *11*(5), 811–. https://doi.org/10.3390/diagnostics11050811

Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., & Houlsby, N. (2020). Big transfer (BiT): general visual representation learning. In: Vedaldi, A., Bischof, H., Brox, T., & Frahm, J. M. (Eds.) *Computer Vision – European Conference on Computer Vision 2020. Lecture Notes in Computer Science*, *12350*, 491-507. https://doi.org/10.1007/978-3-030-58558-7_29

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6),84–90. https://doi.org/10.1145/3065386

Majumder, S., & Ullah, M. A. (2019). Feature extraction from dermoscopy images for melanoma diagnosis. *SN Applied Sciences, 1*(7), 1–11. https://doi.org/10.1007/s42452-019-0786-8

Marcelino, P. (2018). *Transfer learning from pre-trained models*. Retrieved from: https://towardsdatascience.com/transfer-learning-from-pre-trained-models-f2393f124751

Maron, R. C., Haggenmüller, S., von Kalle, C., Utikal, J. S., Meier, F., Gellrich, F. F., Hauschild, A., French, L. E., Schlaak, M., Ghoreschi, K., Kutzner, H., Heppt, M. V., Haferkamp, S., Sondermann, W., Schadendorf, D., Schilling, B., Hekler, A., Krieghoff-Henning, E., Kather, J. N., … Brinker, T. J. (2021). Robustness of convolutional neural networks in recognition of pigmented skin lesions. *European Journal of Cancer (1990)*, *145*, 81–91. https://doi.org/10.1016/j.ejca.2020.11.020

Masood, A., & Al-Jumaily, A. A. (2013). Computer aided diagnostic support system for skin cancer: A review of techniques and algorithms. *International Journal of Biomedical Imaging*, *2013*, 323268–22. https://doi.org/10.1155/2013/323268

Pacheco, A. G. ., Lima, G. R., Salomão, A. S., Krohling, B., Biral, I. P., de Angelo, G. G., Alves Jr, F. C. ., Esgario, J. G. ., Simora, A. C., Castro, P. B. ., Rodrigues, F. B., Frasson, P. H. ., Krohling, R. A., Knidel, H., Santos, M. C. ., do Espírito Santo, R. B., Macedo, T. L. S. ., Canuto, T. R. ., & de Barros, L. F. (2020). PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. *Data in Brief*, *32*, 106221–106221. https://doi.org/10.1016/j.dib.2020.106221

Qureshi, A.S., & Roos, T. (2021). Transfer learning with ensembles of deep neural networks for skin cancer detection in imbalanced data sets. *Arxiv e-prints*, arXiv:2103.12068. https://arxiv.org/abs/2103.12068

Saba, T., Khan, M. A., Rehman, A., & Marie-Sainte, S. L. (2019). Region extraction and classification of skin cancer: A heterogeneous framework of deep CNN features fusion and reduction. *Journal of medical systems*, *43*(9), 289. https://doi.org/10.1007/s10916-019-1413-3

Satheesha, T. Y., Satyanarayana, D., Prasad, M. N. G., & Dhruve K.D (2017). Melanoma is skin deep: a 3d reconstruction technique for computerized dermoscopic skin lesion classification. *Journal of Translational Engineering in Health and Medicine*, *5*, 1-17. https://doi.org/10.1109/JTEHM.2017.2648797

Sharma, N., Jain, V., Mishra, A. (2018). An analysis of convolutional neural networks for image classification: International conference on computational intelligence and data science (ICCIDIS 2018). *Procedia Computer Science*. *132*, 377-384. https://doi.org/10.1016/j.procs.2018.05.198

Shawon, M., Abedin, K. F., Majumder, A., Mahmud, A., & Mishu, M. M. C. (2021). Identification of risk of occurring skin cancer (Melanoma) using convolutional neural network (CNN). *AIUB Journal of Science and Engineering (AJSE)*, *20*(2), 47 - 51. Retrieved from http://ajse.aiub.edu/index.php/ajse/article/view/140

Shi, X., Su, H., Xing, F., Liang, Y., Qu, G. & Yang, L. (2020). Graph temporal ensembling based semi-supervised convolutional neural network with noisy labels for histopathology image analysis. *Medical Image Analysis,60* https://doi.org/10.1016/j.media.2019.101624

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *Arxiv e-prints*, arXiv:1409.1556. https://arxiv.org/abs/1409.1556

Singh, R.D., Mittal, A. & Bhatia, R.K. (2018). 3D convolutional neural network for object recognition: a review. *Multimedia Tools and Applications, 78*, 15951–15995. https://doi.org/10.1007/s11042-018-6912-6

Smith, L. N. (2015). Cyclical learning rates for training neural networks. *Arxiv e-prints*, arxiv:1506.01186.

Staples, M. P., Elwood, M., Burton, R. C., Williams, J. L., Marks, R., & Giles, G. G. (2006).Non-melanoma skin cancer in Australia: the 2002 national survey and trends since 1985. *The Medical Journal of Australia*, *184*(1), 6–10. https://doi.org/10.5694/j.1326-5377.2006.tb00086.x

Subramanian, R. R., Achuth, D., Kumar, P. S., Reddy, K.N.K, Amara, S. & Chowdary, A. S. (2021). Skin cancer classification using convolutional neural networks. *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 13-19. doi: 10.1109/Confluence51648.2021.9377155

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. & Rabinovich, A. (2014). Going deeper with convolutions, *Arxiv e-prints*, arxiv:1409.4842

Tajbakhsh,N., Shin,J.Y., Gurudu,S.R., Hurst,R.T., Kendall,C.B, Gotway,M.B. & Liang, J. (2016) Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Trans Med Imaging, 35*(5):1299-1312. doi: 10.1109/TMI.2016.2535302

Tanveer, S. M., Khan, M. U. K., & Kyung, C-M. (2020). Fine-tuning DARTS for image classification. *Arxiv e-prints*, arXiv:2006.09042.

Tschandl P. (2018). Sequential digital dermatoscopic imaging of patients with multiple atypical nevi. *Dermatology, Practical & Conceptual*,*8*(3):231-237. doi: https://doi.org/10.5826/dpc.0803a16

Tschandl, P., Rosendahl, C. & Kittler, H. (2018). The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data, 5*(180161). https://doi.org/10.1038/sdata.2018.161

Yu, F. & Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions, *Arxiv e-prints*, arxiv:1511.07122.