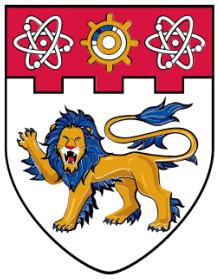


NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

BC3409 AI in Accounting and Finance Group Project

TrueHealth, the Future of Melanoma Detection

AY22/23 Sem 1

Team Members:
Sng Yi Xuan (U2021009B)
Sydney Teo Wen Xuen (U2021555B)
Ivan Lua Yangzhi (U2022506F)
Cao Thuc Anh (U2010720G)
Teresa Zhang Han Yu (U2022886C)
Yam Hui Jing (U2021656A)

Class/Team:
S02 /Group 3

Table of Contents

1. Executive Summary

Melanoma is the most invasive skin cancer with the highest risk of death. While it's a serious skin cancer, it's highly curable if caught early. Therefore, the early detection of Melanoma can vastly increase chances for cure, and is critical for a patient's survival.

However, there are limitations to the current methods for early detection of Melanoma. The current practice of initial melanoma diagnosis has pretty low average diagnostic accuracy due to the limitations to dermoscopy in recognising melanoma as well as the severe shortage in trained expertise. The public also has no reliable, affordable and convenient screening methods for them to diagnose melanoma early other than looking out for warning signs.

As such, this report's main objective is to use machine learning to address the limitations and complement the current methods for the early detection of Melanoma. We have two target audiences, the healthcare industry and potential patients (general public), as we feel that this issue concerns them the most. Our area of focus is Australia as she has the highest number of Melanoma cases.

Since early detection for Melanoma is diagnosed visually, we used Convolutional Neural Network (CNN), a class of deep neural networks, most commonly applied to analyse visual imagery, to differentiate images of moles.

We will first explore CNN using TensorFlow for detecting malignant and benign melanoma. We will be using the default parameters as learnt in class, and analyse its shortcomings. From there, we will finetune its hyperparameters for further improvement of the model. On top of that, we will also use popular CNN models such as ResNet50, AlexNet. We will also use Inceptionv3 (Orange). Finally, we will compare the different CNN models and evaluate the most efficient model for our problem statement.

From our results, we observe that AlexNet-Based has the highest accuracy. ResNet50, on the other hand, has the lowest FNR and highest recall. As previously mentioned, in the context of healthcare, it is extremely important to have low false negative rates and our chosen model should be able to detect almost all positive cases of Melanoma so that no patients that have a cancerous melanoma are left undetected and untreated. Hence, a low FNR and high recall is preferred, in which case ResNet50 is the most optimal choice. Therefore, our group has decided to use **ResNet50 as our chosen model** for our application.

For this project, we have developed **TrueHealth**, a machine learning web application hosted on Heroku that detects and diagnoses the severity of Melanoma. Our application consists of our **ResNet50** model, hosted on our Flask Backend to provide diagnosis predictions based on user-uploaded images. Our application is designed in such a way that addresses the concerns for both the potential patients and the healthcare professionals. The general public is able to use the application for the purpose of self-diagnosis. Furthermore, dermatologists can use our application to get a second opinion or expedite the diagnosis process, thereby increasing the probability of a correct diagnosis and the expedition of the treatment process.

2. Problem Statement

Melanoma is a type of skin cancer that evolves from the rapid growth of melanin-producing cells, Melanocytes, which are located in the skin's epidermis. It can be malignant or benign (Skin Cancer (Non-Melanoma) - Introduction, 2022). A cancerous tumour is malignant, meaning it can grow and spread to other parts of the body. A benign tumour means the tumour can grow but will not spread.

According to study, melanoma is the 17th most common cancer worldwide. In 2020, an estimated 57,043 people worldwide died from melanoma. However, Melanoma has a 99% survival rate if it is caught before spreading to the lymph nodes (Melanoma - Statistics, 2022). Once it spreads to nearby nodes, though, that rate drops to 66%. For treatment wise, surgery can be performed to remove the tumour throughout the stages of melanoma. Excisional biopsy and skin grafting will no longer be sufficient to treat the disease if it has been spread to the lymph nodes or other parts of the body (Melanoma Treatment (PDQ®)–Patient Version, 2022). In short, not recognising a melanoma when it is present (a false-negative test result), delays surgery to remove it, risking cancer spreading to other organs in the body, and possibly death (Cochrane, n.d.).

With that, we researched and found out that melanoma is more common in Whites than in Blacks and Asians (NCBI - WWW Error Blocked Diagnostic, n.d.). Studies also indicate that melanoma mortality rate is 5 times higher in developed countries (Dermatologytimes, 2020). Australia, being the country with the highest overall rate of melanoma of 16,171 in 2020, becomes the focus of our project (WCRF International, 2022).

Melanoma diagnosis and treatment varied according to the severity or stage of disease. There are only two ways to detect Melanoma thus far, 1) through the naked eye, or 2) through a screening test performed by medical professionals – epiluminescence microscopy, or dermoscopy. Thus, the following are the two main stakeholders whom we observe the problem arises from:

(a) Healthcare industry

Firstly, the current practice of initial melanoma diagnosis is clinical and subjective, relying mainly on the use of naked eye examination (Thanh-Toan et al., 2018). Identifying a potential skin cancer using naked is not easy as melanomas come in many forms and may display none of the typical warning signs. It has an average diagnostic accuracy of only about 85%, even when it is performed by trained expertise. Melanoma can also be detected through a painless medical technique where doctors can evaluate the patterns of size, shape, and pigmentation in pigmented skin lesions (Melanoma - Screening, 2022). However, there are limitations to dermoscopy in recognising melanoma, reducing the diagnostic accuracy for melanoma (NCBI - WWW Error Blocked Diagnostic, n.d.-b).

To worsen the situation, there is a severe shortage of dermatologists in Australia (Somerville, 2021). There are just 550 practicing dermatologists in Australia, and the Department of Health predicts the specialist workforce will increase to a shortage of 90 full-time equivalent

dermatologists by 2030. That is almost 15 per cent less than what is required to meet the needs of the population. This is critical to Australia as the diagnostic accuracy is highly dependent on the trained expertise of dermatologists, which is what it is lack of.

(b) Public

First, the public currently has no means to conduct a reliable self diagnosis of melanoma. This greatly diminishes the chances for patients to monitor independently for early detection. Melanoma in Australia is mainly detected by chance, by either the patient presenting for a routine skin check or with a lesion of concern, or by the doctor detecting a lesion incidentally.

Second, there is a lack of affordable screening available for patients. In Australia, the mean first-year costs of melanoma per patient ranged from AU$644 (95%UI: $642, $647) for melanoma in situ to AU$100,725 (95%UI: $84,288, $119,070) for unresectable stage III/IV disease (National Library of Medicine, 2022). Skin cancer represents the most expensive cancer for Australia's health system.

Through research and evaluation of skin clinics in Australia, we found that skin treatments range as such (Molecheck, 2022):

1. Minor biopsy: $90 - 120
2. Standard/ excision biopsy: $160 - 240
3. Full body skin cancer check: $120
4. Full body skin cancer check & mole mapping: $170

These biopsies are performed by specialised skin cancer doctors who examine the patients' skin with a dermatoscope. They will then further examine suspicious spots by examining photographs under high magnification.

Third, the current diagnosis procedure for melanoma is tedious and long. Currently, patients have to book an appointment to even do a screening evaluation of their melanoma condition by health professionals. Patients who visit General Practitioners (GPs) will have their skin photographs taken and sent to skin specialists. If the melanoma images are evaluated to be suspicious, the skin specialist will have to remove the mole and send it for testing (biopsy) to check whether it is malignant. A biopsy is usually done using local anaesthetic. All in all, it takes around 2-3 weeks (Cancer Research, 2022). to get the biopsy results. The patients might also require further treatment if there are any cancer cells left behind. These would result in high healthcare and medical costs.

3. Opportunities

With these existing problems in mind, we analyse the current market of telehealth to see whether it is a profitable market for us to venture into.

3.1 Market analysis

The rising penetration of the internet and innovation in smartphones is enabling to address the gaps in the delivery and availing of telehealth services conveniently. Moreover, the demand for telehealth services witnessed tremendous growth over the past year due to the COVID-19 pandemic and the restrictions imposed in order to curb the infection. COVID-19 hindered the delivery of healthcare services which enabled most healthcare facilities to shift from traditional to virtual care methods. Further, the growing need to monitor health and wellness in order to manage chronic diseases virtually is driving the market growth (Grand View Research, n.d.).

The global telehealth market was estimated at USD 41 billion in 2021 and expected to reach USD 224.8 billion by 2030 and poised to grow at a compound annual growth rate (CAGR) of 18.8% during the forecast period 2022 to 2030 (Precedence Research, n.d.) (Figure 3.1).

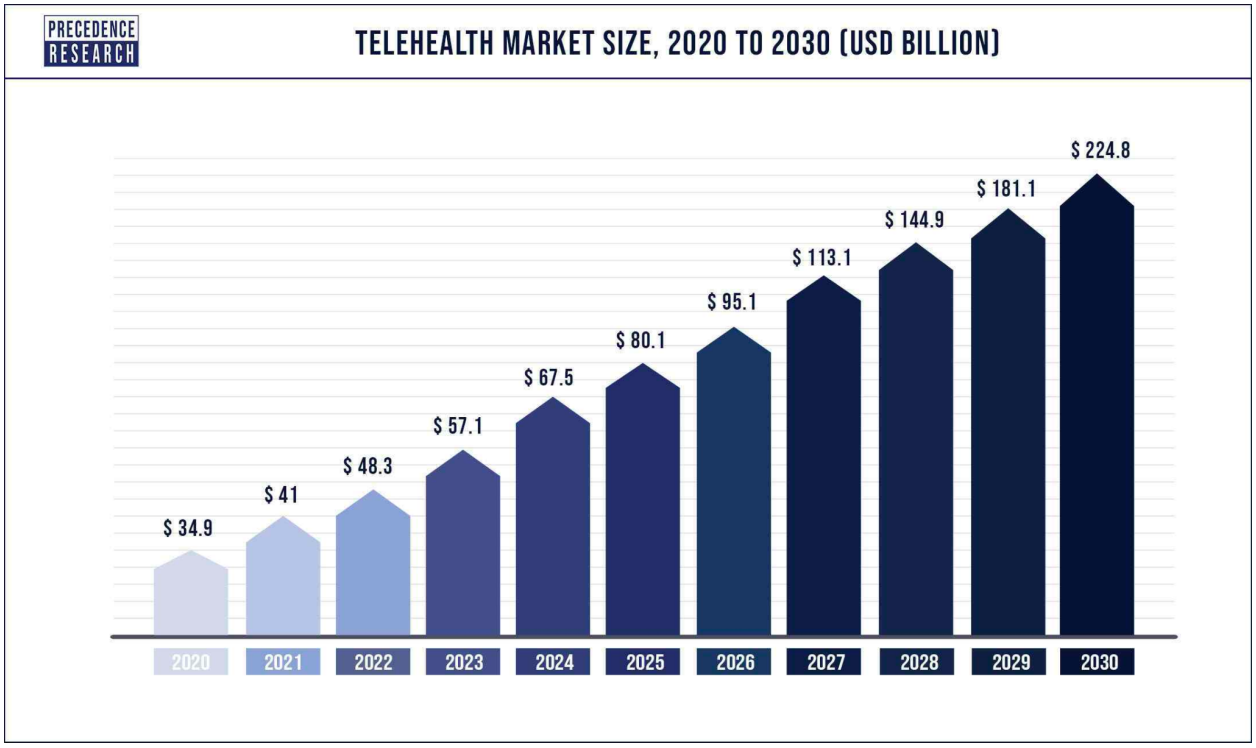


Figure 3.1: Global Telehealth Market Size (USD Billion) from 2020 to 2030

The market size, measured by revenue, of the Telehealth industry in Australia is $103.1 million and has grown 26.4% per year on average between 2017 and 2022 ((IBISWorld - Industry Market Research, Reports, and Statistics, n.d.)) (Figure 3.2).

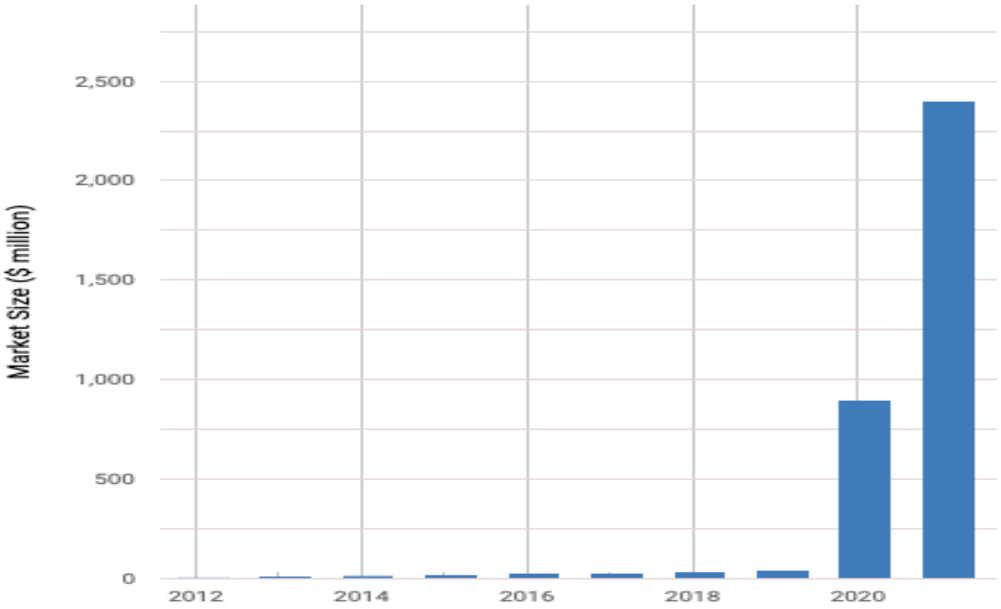


Figure 3.2: Australia Telehealth Market Size (USD Billion) from 2012 to 2021

Thus, our team feels that this market will be profitable and we will grasp this opportunity and enter the growing market.

3.2 Competitor analysis

As healthcare technology evolves, many organisations have thrived to propose state-of-the-art solutions. A review of skin cancer apps available in 2019 found there were 66 apps commercially downloadable for consumers, often offering multiple functionalities, with just under half (49%) aimed at supporting monitoring and tracking of lesions, followed by artificial intelligence image lesion analysis (39%), education provision (38%) and teledermatology services (27%) (Janda M et al, 2022). In the scope of this proposal, direct and notable competitors to be analysed are Miiskin, MoleScope and SkinVision.

We will be looking at the features of the respective products, its price, its success (by looking at the number of downloads), the market it is targeting, and lastly, if it is able to diagnose the user, if yes what is its accuracy.

	MiiSkin	MoleScope	SkinVision
Diagnosis?	No	No	Yes Accuracy: 95%
Features	Skin check/tracking app	Hardware skin magnifier (imaging, archiving, communication)	Self-examine, regulated medical service

Price	• B2B: $300 - 600/month • B2C: $2.31 - 5.28/month	• MoleScope Lite: $49.99 • MoleScope II: $299	• Single check: $6.99 • 3-month plan: $29.99 • 12-month unlimited check: $49.99
Downloads	100K+ (Play Store)	10K+ (Play Store)	500K+ (Play Store)
Markets	Europe, America	Global	Global
Unique selling proposition	Mole Sizing (AR)	Portable skin view	Able to diagnose at a high accuracy

Table 3.1: Differences between competitors

Generally, all the competitors charge a certain amount to utilise premium features such as supervising more than three moles, or checking of moles for more than once. From the analysis above, SkinVision is our closest competitor since it is also using Artificial Intelligence. It has a high accuracy and has a large consumer base with a strong brand image. This motivates us to take advantage of advanced technology to create a free-of-charge application that is accessible to even those of low-income.

4. Project Objectives and Goals

As such, we need to provide an economical, easy-to-use technology-driven solution that can help people check skin anomalies at their convenience with just a single click on their phones, which facilitate possible early treatment and create a healthy community.

Our project's main objective is to provide an additional way for early detection of Melanoma using Convolution Neural Network (CNN), such that it addresses the limitations of the two methods to detect Melanoma and compliments them, to overall reduce the mortality rate for potential patients with Melanoma.

4.1 Business-to-Business

Unfortunately, even with dermoscopy, some melanomas remain difficult to diagnose (NCBI - WWW Error Blocked Diagnostic, n.d.-c). However, these difficult to diagnose melanomas often reveal subtle clues that allow for their correct identification. Thus, we aim to train a CNN model that will be able to identify these clues, complement existing tools in the diagnosis of melanoma and assist medical personnels. For the front-end, we aim to make use of tools and features such as brightness, zoom, and annotations that can help the medical personnels.

4.2 Business-to-Consumer

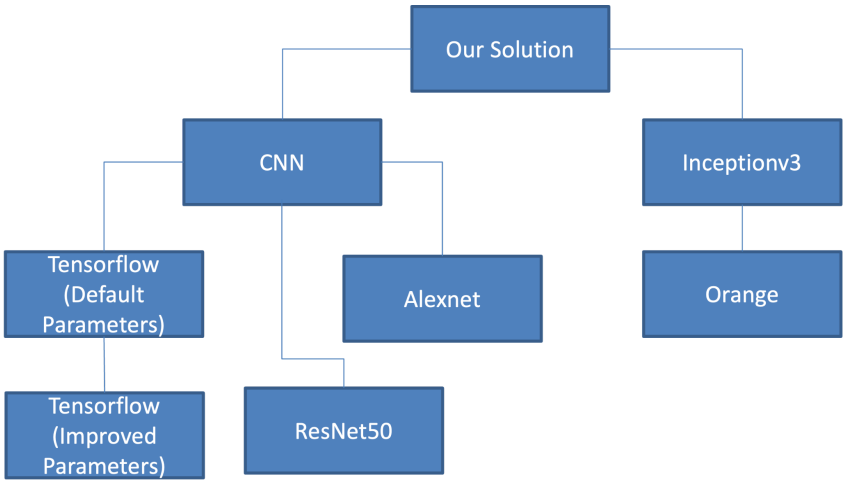
As mentioned previously, the general public has no means to detect melanoma early, and can only rely on their naked eyes to look out for the warning sign of Melanoma, which is not easy to tell and might be inaccurate. Thus, to help them with early self-detection and analyse the

severity before doctor examination, we aim to design a web-application that is free and highly accessible to the public, and find the best CNN model to ensure the accuracy of the application.

5. Our Solution

To build our application, we first have to build an accurate and reliable machine learning model that can predict the severity of melanoma.

We will first explore CNN using TensorFlow for detecting malignant and benign melanoma. We will be using the default parameters as learnt in class, and analyse its shortcomings. From there, we will finetune its hyperparameters for further improvement of the model. On top of that, we will also use popular CNN models such as ResNet50, AlexNet. We will also use Inceptionv3 (Orange). Finally, we will compare the different CNN models and evaluate the most efficient model for our problem statement.



Overview of approach

5.1 Data Visualization

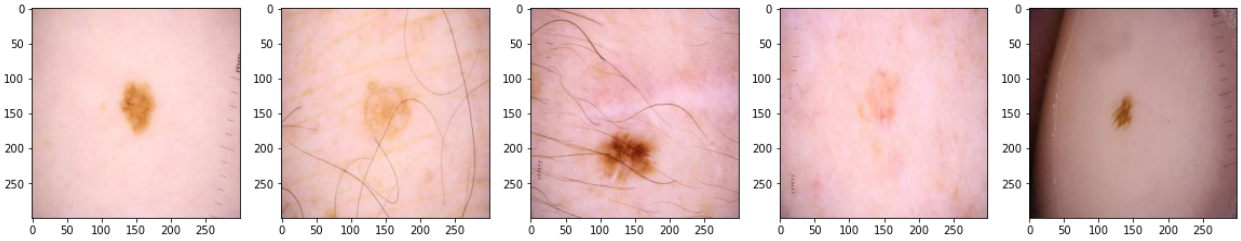


Figure 5.1.1: Benign samples

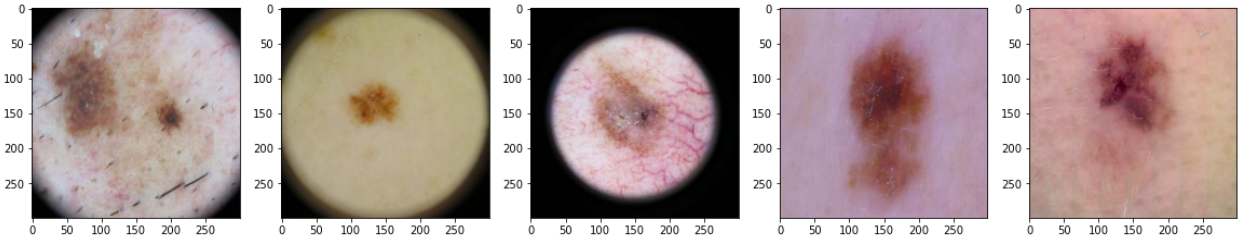


Figure 4.1.2: Malignant samples

From the images above, we can see that a malignant melanoma and a benign one may look similar. Although some malignant melanoma is easy to tell from its distinct features such as being abnormally discoloured or shaped, some benign melanoma can also appear brown and patchy. An untrained eye is highly unlikely to be able to distinguish between a benign case and malignant case accurately consistently. Therefore, CNN deep learning can serve as a beneficial tool for a quick and effective screening indicator.

Dataset

	Train	Test
Malignant	5000	500
Benign	4605	500

Table 5.1.1: Number of images in dataset

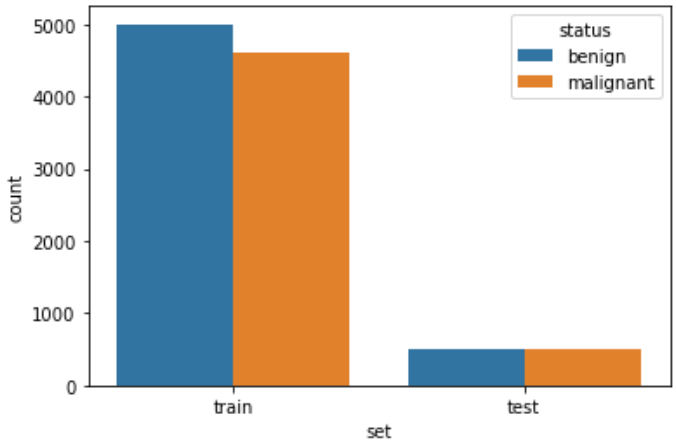


Figure 5.1.3: Barplot count of benign and malignant melanoma in dataset

5.2 Initial Approach using TensorFlow

Introduction of our machine learning model

Convolutional Neural Network (CNN) is a type of Deep Neural Network because it has a high network depth and has been widely used to detect and classify objects in an image. Hence, we will use CNN for the detection of melanoma skin disease. The CNN architecture consists of 4 key layers:

Layer	Description
Convolution Layer	The convolution layer is the first layer where the input image is convolved using filters. Every image is considered as a matrix of pixel values. Using a sliding window format, filters are applied across the input image to perform a weighted sum calculation. This process continues until the convolution operation is complete.
ReLU (Rectified linear unit) activation layer	In the second layer, the ReLU activation function is used to perform element-wise operations and set negative pixels to 0. This introduces non-linearity to the network, and the generated output is a rectified feature map.
Pooling Layer	Pooling is a down-sampling operation that reduces the dimensionality of the features extracted from the previous layers. The pooling layer uses various filters to identify different distinct features of the image like edges, corners, eyes, and beak
Fully Connected Layer	In the last layer, flattening is done where the previously pooled features are transformed into a single matrix. This is fed as input to the neural network for weight and bias updates. The final activation function used will be the sigmoid function as we are only predicting binary results, whether there is benign or malignant melanoma on the skin

Table 5.2.1: Types of layers in CNN architecture

Methods to implement

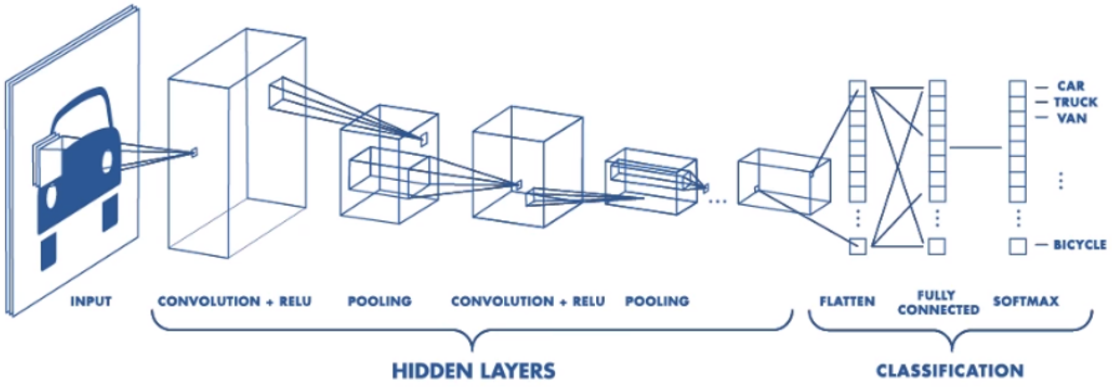


Figure 5.2.1: CNN Layers

For CNN in Keras, we only need to modify the network structure and input format. We decided to use what was taught in the seminar as the default parameters. We will use 4 convolution layers. The input shape will be (150, 150, 3), where the resolution of the image used is changed to 150x150 pixels and 3 represents that we are using RGB for the image. The image is convoluted using 3×3 filters on each hidden layer with the number of output channels on each layer is 32, 32, 64, 64 respectively. Maxpool is set to (2,2) with strides 2, to reduce the size of the image. We will then use default parameters of flatten. Lastly, softmax activation function helps classify the condition of skin image into 2 classes namely, malignant or benign.

Evaluation of initial approach

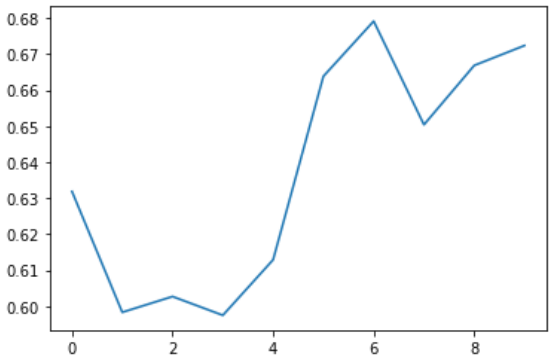


Figure 5.2.2: Model Accuracy

The model accuracy diagram plots the accuracy on each epoch. After running the model for 10 epochs with 4 Convolution Layers, it was observed that the model accuracy peaked around 0.68 at the 6th epoch. There was a sharp increase from the 4th to the 6th epoch, indicating the network is learning fast. A 0.6724 accuracy shows that the result is poor. We stopped at 10 epochs as there was no trailing tail. Therefore, it is unlikely that increasing the number of epochs will improve the model accuracy further. This suggests that we may need to change the configuration of the model parameters and augment the data more appropriately.

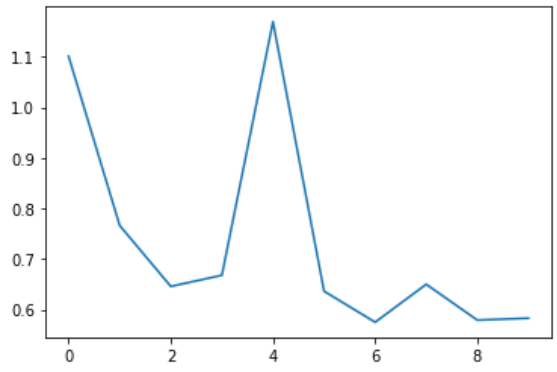


Figure 5.2.3: Model Loss

Loss is the penalty for a bad prediction. If the model's prediction is perfect, the loss is zero, else the loss is greater. The model loss curve has a general decreasing trend, albeit with fluctuations, indicating the training is unstable. Hence, we may need to adjust parameters such as learning rate to ensure the model loss diagram converges. An ideal model loss diagram would have a smooth decreasing trend, with little to no fluctuations.

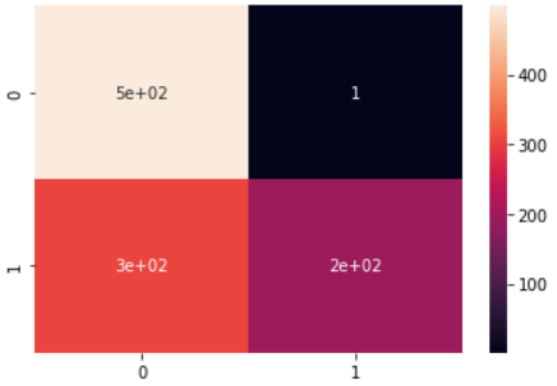


Figure 5.2.4: Confusion Matrix for Tensorflow (Default parameters)

From the confusion matrix, we can obtain metrics such as accuracy as it can be calculated as a ratio of the total number of correct predictions to the total number of predictions generated by the model. Here, we obtained a model accuracy of 0.694, which is low and the false negative rate (FNR) is 0.61. This is high as the model incorrectly predicts that a melanoma image is benign when it is actually malignant 61% of the time.

Lastly, we analyse the precision and recall results which can also be calculated from the confusion matrix. Precision is the proportion of positive identifications that was actually correct. We obtained a precision of 0.9948 (4dp), meaning that when the tensorflow model predicts the melanoma image is malignant, it is correct 99.48% of the time, which is relatively high and desirable. Recall is the proportion of actual positives that was identified correctly. The model has a recall of 0.39, meaning that it correctly identifies 39% of all malignant melanoma images.

	Accuracy	False Negative Rate	Precision	Recall
Result	0.694	0.610	0.995	0.390

Figure 5.2.5: Summary of Results for Default Model

Overall, although the precision is high, we find that the model accuracy, FNR, and recall can be further improved and that the model is definitely not suitable to be deployed for public use.

5.2 Improvement to Initial TensorFlow Model

As observed in the previous section, the accuracy of our TensorFlow default model is only 0.694 with a high false negative rate of 0.610. In this section, we will aim to improve the accuracy of our model by performing data augmentation as well as by tuning the hyperparameters of the our TensorFlow model.

Data Augmentation

The accuracy of deep learning models largely depends on the quality, quantity, and contextual meaning of training data. However, data scarcity is one of the most common challenges in building deep learning models. One low-cost and effective method—data augmentation helps to reduce dependency on the collection and preparation of training examples and build high-precision AI models quicker. Data augmentation is the addition of new data artificially derived from existing training data. Techniques include resizing, flipping, rotating, cropping, padding, etc. It helps to address issues like overfitting and data scarcity, and it makes the model robust with better performance. We look at 4 different data augmentation techniques:

1. **Rescalling:** Scaling provides more diversity in the training data of a machine learning model. Scaling the image will ensure that the object is recognized by the network regardless of how zoomed in or out the image is. [1. / 255]
2. **Zoom Range**: Zoom augmentation randomly zooms the image in and either adds new pixel values around the image or interpolates pixel values respectively [zoom_range=0.2]
3. **Rotation:** Rotation helps the models not consider the angles of an image to be a distinct feature for prediction [rotation_range=15]
4. **Flip:** By Flipping images, the optimizer will not become biased that particular features of an image are only on one side [horizontal_flip=True]

Deciding the Best Parameters

A deep learning CNN consists of three layers: a convolutional layer, a pooling layer and a fully connected (FC) layer. To improve the TensorFlow model performance, we can tune the parameters in each of these layers. In this section below, we will explain the reasons for our choice of parameters:

1. Convolutional Layer

```
nb_filter = 32
image_size = (300,300,3)

model = Sequential()
model.add(Conv2D(nb_filter,(3,3), input_shape=image_size, activation = 'relu'))
model.add(MaxPool2D(pool_size=(2,2),strides = 2))

model.add(Conv2D(nb_filter*2, (3, 3), activation = 'relu'))
model.add(MaxPooling2D(pool_size=(2, 2),strides = 2))

model.add(Conv2D(nb_filter*4, (3, 3), activation = 'relu'))
model.add(MaxPooling2D(pool_size=(2, 2),strides = 2))

model.summary()
```

Figure 5.2.1: Embedding Layer Code

Firstly, for the convolutional layer, we have decided to use ***3 layers*** instead of the usual 4 because from our iterations, we experience that in general “more convulational layers the better”. However, from our experience, after about two or three layers, the accuracy gain becomes rather small. Since we will be training the model on ***100 epochs***, we will decrease the complexity of the model by decreasing the number of layers so as to decrease the training time.

Secondly, **image size**. The optimal image size depends on the data given, we should use the image size of the images that we are given. Downscaling images will cause bigger images to be down scaled, this makes it harder for CNN to learn the features required for classification or detection as the number of pixels where the vital feature will be present is significantly reduced. On the other hand, upscaling will cause small images to be upscaled and padded with zero, then NN has to learn that the padded portion that has no impact on classification. Additionally, larger images are also slower to train and might require more VRAM. Therefore, it is important that we use the ideal image size to obtain the best results for training (Ramalingam, 2021).

As such, we perform some exploratory analysis on the image sizes and from the figures below, we see that for all the images (test and train/ benign and malignant), the image size is (300, 300). Therefore, for our CNN model, we will use ***input_shape = (300,300,3).***

```
Train Benign Dataset
Total Nr of Images in the dataset: 5000
```

	FileName	Size	Width	Height	Aspect Ratio
0	melanoma_0.jpg	(300, 300)	300	300	1.0
1	melanoma_1.jpg	(300, 300)	300	300	1.0
2	melanoma_10.jpg	(300, 300)	300	300	1.0
3	melanoma_100.jpg	(300, 300)	300	300	1.0
4	melanoma_1000.jpg	(300, 300)	300	300	1.0

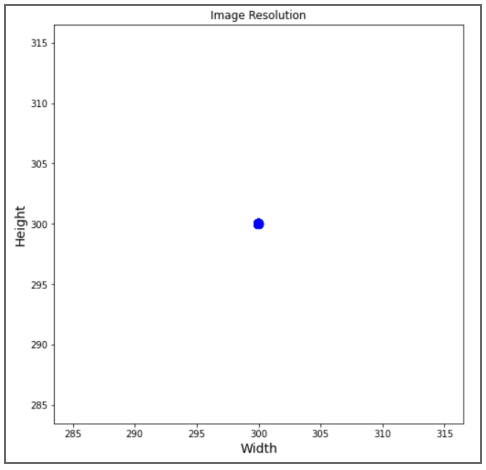


Figure 5.2.2: Image sizes of Melanoma images

Next, we look at the **ideal number of filters** and the **filter size**. Every layer of filters is there to capture patterns. For example, the first layer of filters captures patterns like edges, corners, dots etc. Subsequent layers combine those patterns to make bigger patterns (like combining edges to make squares, circles, etc.) Therefore, we have decided to start with **32 filters** as default and as we move forward in the layers, we will increase the filter size by a **multiple of 2** each time. This is because the patterns will get more complex as we move forward in the layers and hence there are larger combinations of patterns to capture. For filter size, we will also maintain the **default size of 3x3.** Lastly, we will not be adding any padding to the images as from our observations, most of the moles on the melanoma images appear in the centre of the image and thus not appear on the sides or edges of the image. Therefore, padding will be unnecessary.

2. Pooling Layer

Moving onto the pooling layer, we will use MaxPool of **size 2x2** with a **stride of 2**. A problem with the output feature maps from the convolutional layer is that they are sensitive to the location of the features in the input. One approach to address this sensitivity is to down sample the feature maps. This has the effect of making the resulting down sampled feature maps more robust to changes in the position of the feature in the image, referred to by the technical phrase “local translation invariance.” By default, the size of the pooling operation is almost always 2x2 pixels. We use Max Pooling instead of Average Pooling because it has the added functionality of noise-suppressing, as it works on discarding those activations which contain noisy activation.

3. Fully Connected (FC) Layer

```
nb_fc_neurons = 512

model.add(Flatten())
model.add(Dense(nb_fc_neurons))
model.add(Activation('relu'))
model.add(Dropout(0.5))

model.add(Dense(1))
model.add(Activation('sigmoid'))
```

```
model.compile(loss='binary_crossentropy',
                  optimizer='adam',
                  metrics=['accuracy'])
```

Figure 5.2.3: Fully Connected Layer Code

Next, we come to the last layer, FC. The FC layer is where image classification happens in the CNN based on the features extracted in the previous layers. The first step taken in the FC layer is to flatten the 3D-input into a 1-D array which can then be fitted into the our neural network model. We have decided to use **512 neurons** with **1 hidden layer**. (Malik, 2019) There is a consensus that the performance difference from adding additional hidden layers: the situations in which performance improves with a second (or third, etc.) hidden layer are very few. One hidden layer is sufficient for the large majority of problems. From our testing, we can concur with the research above. We had previously ran the model using 3 hidden layers with 512 - 128 - 64 neurons, but from our experience, it resulted in lower accuracy as well as longer training time. Therefore, for faster training and better accuracy, we will use 1 hidden layer with 128 neurons. Additionally, we also use a **dropout of 0.5** to prevent overfitting.

The output of our fully connected layer is 1 node that is a binary classification. This is because our Melanoma dataset predicts whether an image is benign (0) or malignant (1). Because the output is binary, we will use the **activation function sigmoid** which is suitable for binary classification (Basta, 2020). For multicategorical classification, softmax would be the ideal choice. Additionally, as our output is either 0 or 1, the loss function most appropriate would be **binary crossentropy** and similarly, if the output is multicategorical, the loss function most appropriate would be categorical cross entropy (Admin, 2021).

```
train_data_size = 9605
batch_size = 16

h = model.fit_generator(
        train_generator,
        steps_per_epoch = train_data_size // batch_size,
        epochs= 100,
        validation_data = test_generator,
    )
```

Figure 5.2.4: Model Fit Code

Finally, with the CNN model fully defined, it is time to train the model. Because data augmentation is performed on each epoch, instead of using *model.fit* we will use *model.fit_generator* to fit the training data to the model. We have decided to train on a **batch size of 16** rather than 5. In practical terms, to determine the optimum batch size, we recommend trying smaller batch sizes first, usually 16 or 32 (Kandel & Castelli, 2020). The number of batch sizes should be a power of 2 to take full advantage of the GPUs processing. As a result, we have decided to use a batch size of 16.

Referring back to the default TensorFlow model that was trained on 10 epochs, we observe that the model is underfited at 10 epochs as we see that there is no tail in the accuracy and loss graph, the model can still be further improved by increasing the number of epochs, as a result, for the improved model, we have decided to run the training on 10 epochs to allow the model to model to go from unfitted to optimal. With that, the diagram below summarises the the architecture used to for our improved TensorFlow CNN model.

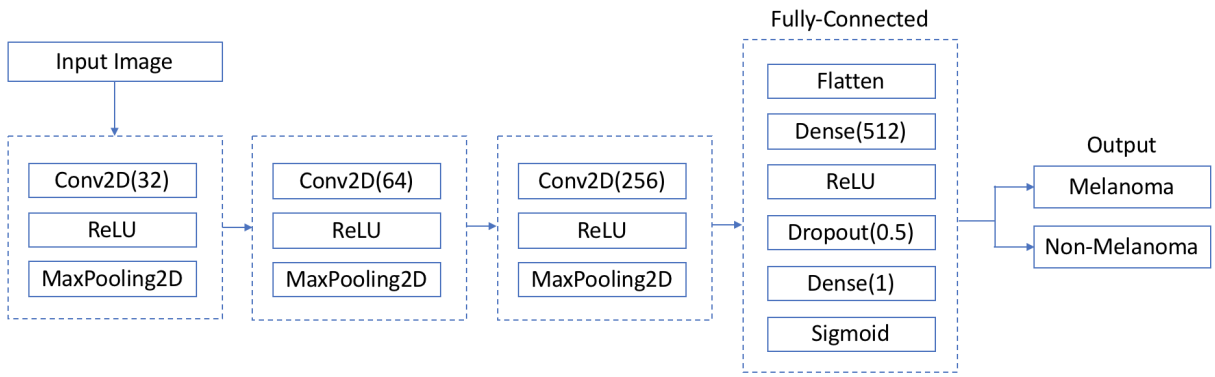


Figure 5.2.5: TensorFlow Architecture Block Diagram

Results Obtained

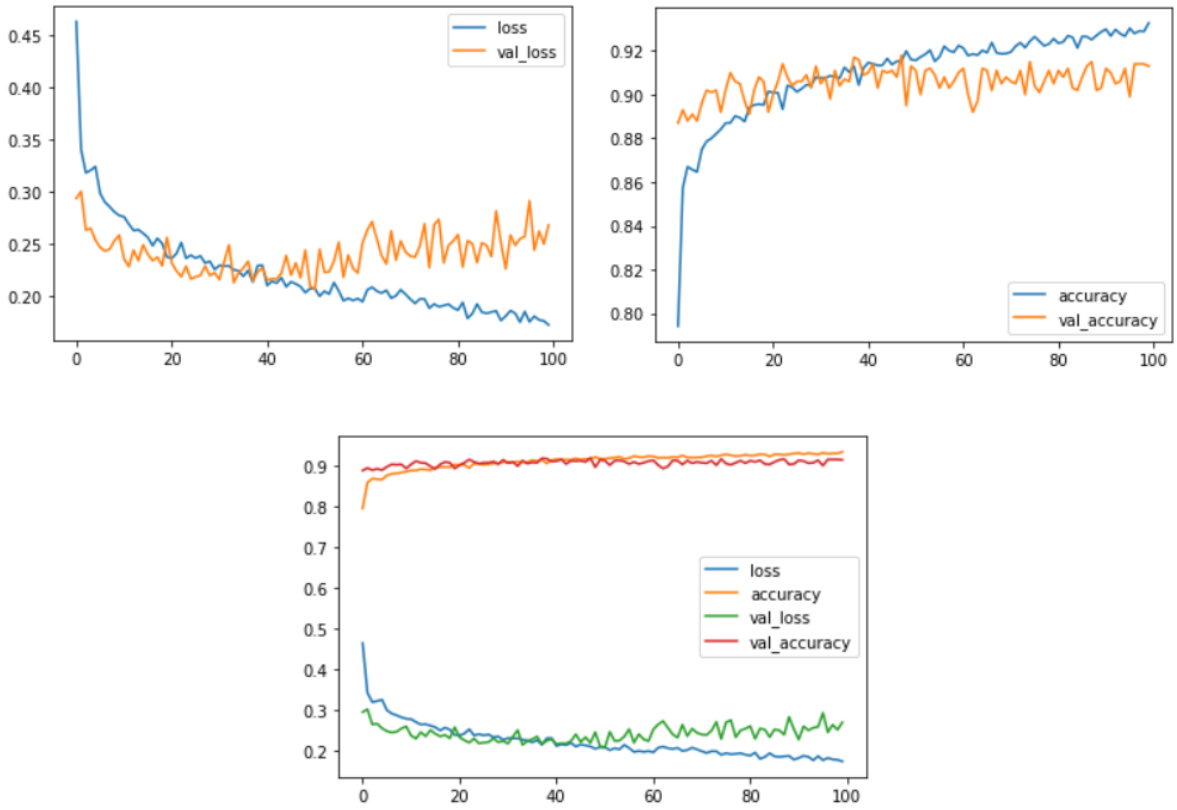


Figure 5.2.6: Accuracy & Loss Graph

Referring to the figure above, we observe for training, the loss curve decreases consistently while the accuracy curve increases consistently where we see a small tail at the ends of both curve suggesting that the model is trained optimally. Comparing it to the validation loss and accuracy curve, we see that the validation loss curve fluctuates and is seen to be increasing slightly. This is not a ideal as optimally, we would want the validation curve to be as similar to that to the training loss curve. Comparing the validation accuracy curve however, we also see that it seems to fluctuate slightly as well and is showing a slight increasing trend as well. We observe that towards the end of the epochs, it seems to flatten thus signifying that the model has reached the optimal number of training epochs and further increasing the number of epochs will not result in better results. Hence, we concluded that 100 epochs is sufficient to train the TensorFlow model.

	Accuracy	False Negative Rate	Precision	Recall
Result	0.913	0.114	0.937	0.886

Figure 5.2.7: Summary of Results for Improved Model

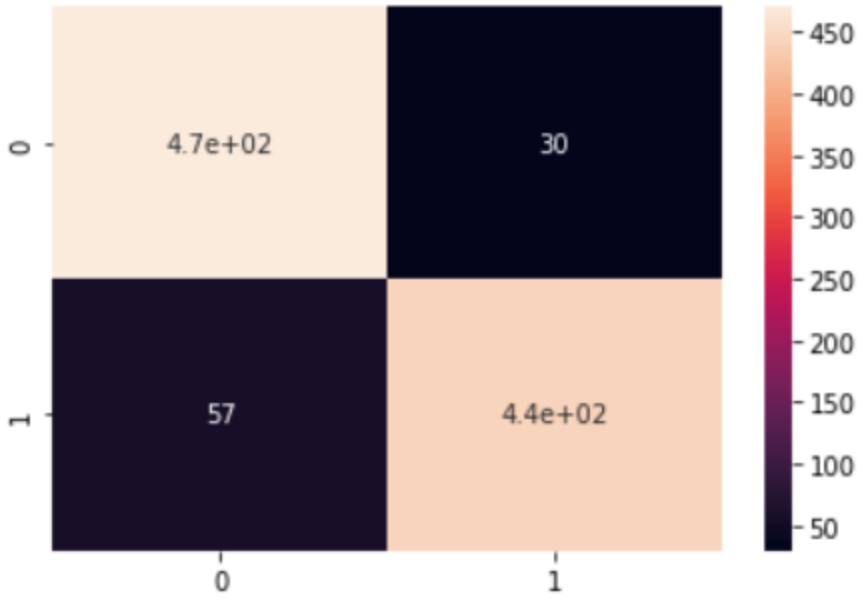


Figure 5.2.8: Confusion Matrix for Improved Model

Evaluating the results obtained from Figure 5.2.7 and the above confusion matrix, we see that our accuracy has increased significantly from 0.694 to 0.912. Additionally, the false negative rate has decreased from 0.610 to only 0.114. This suggests that with the new changes to the model, we saw a huge improvement to the performance. We are now able to achieve higher accuracy and lower FNR which suggests that the model is now better able to predict melanoma classes and is less likely to wrongly predict benign. Evaluating the precision and recall, we see a rather small decrease in precision but a huge increase in recall from 0.390 to 0.886. This means for the improved model, the precision and recall curve will have an AP close to 1 suggesting that the model is optimal and could be potentially used as the model for our application.

5.3 Comparison to Other CNN Models

In addition to TensorFlow model, we have explored Orange, ResNet50 and Alexnet model. The introduction of model as well as methods of implementation are presented below.

5.3.1 Orange

Introduction of Orange

Orange is a powerful platform to perform data analysis and visualisation, see data flow and become more productive (Bioinformatics Laboratory, University of Ljubljana, n.d.-a). It provides a clean, open source platform and the possibility to add further functionality for all fields of science. It performs basic functionalities such as reading the data, showing a data table, selecting features, training predictors, comparing learning algorithms, visualising data elements, etc. The following is the diagram of our workflow in Orange:

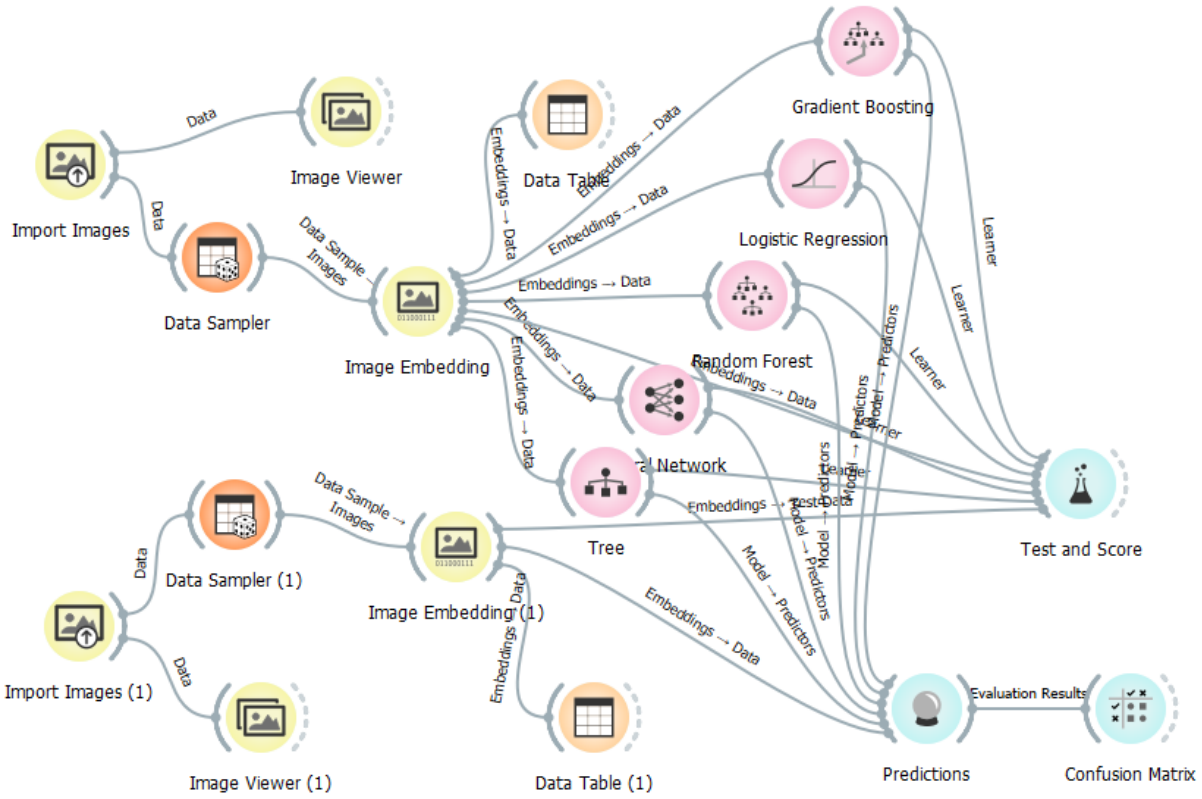


Figure 5.3.1: Workflow in Orange

Methods of implementation

We have 2 widgets for Import Images where we import the files of train and test data respectively into Orange. We can use Image Viewer to display images from the datasets imported previously (Bioinformatics Laboratory, University of Ljubljana, n.d.-b).

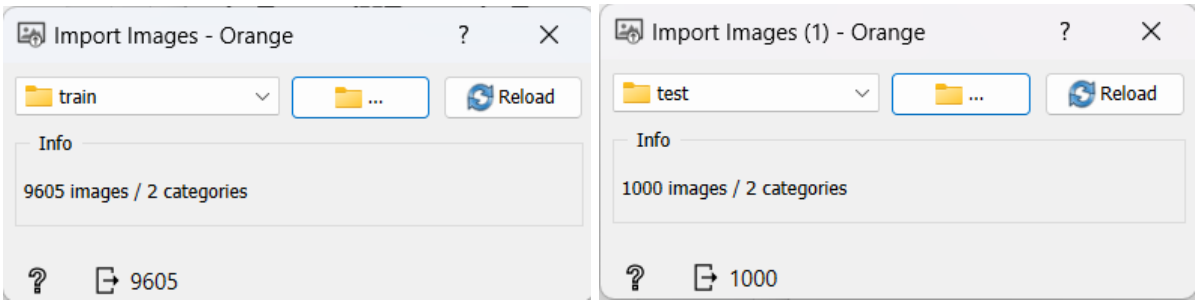


Figure 5.3.1.1: Importing train and test images

From Import Images, we will send a data table containing a column with image paths into Image Embedding. Image embedding is implemented through deep neural networks. It takes in a list of images from data sampler and output embeddings - images which will be represented with a vector of numbers (Bioinformatics Laboratory, University of Ljubljana, n.d.-b). It reads images and uploads them to a remote server or evaluate them locally. We have chosen Inceptionv3, which is Google's deep neural network for image recognition as the embedder.

With image embedding, the dimensionality of input data could be greatly reduced (Blog, 2020). Input images are converted into low-dimensional vectors that can be more easily used by other computer vision tasks. The key to good embedding is to train the model so that similar images are converted to similar vectors. Activations of the penultimate layer of the model, which represents images with vectors are used.

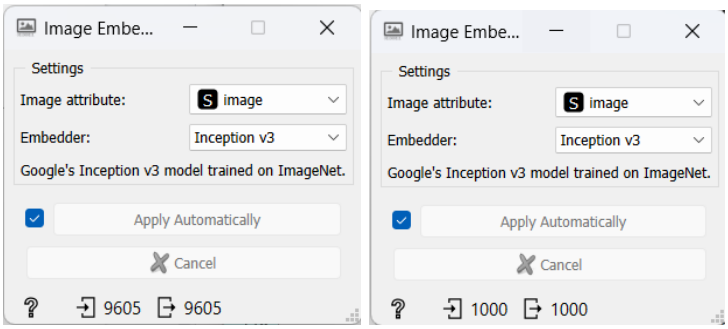


Figure 4.3.1.3: Image embedding

Embedder - InceptionV3

The inception V3 is just the advanced and optimized version of the inception V1 model. It has a total of 42 layers and a lower error rate than its predecessors (Narein T, 2021). The architecture of an Inception v3 network is progressively built, step-by-step, as explained below:

1. Factorized and smaller convolutions

Generous dimension reduction in inception v1 model has been further improved in inception v3 model (Advanced Guide to Inception V3 | Cloud TPU |, n.d.). For example, consider the basic module of inception v1 module in Figure 5.3.1.4. It has a 5×5 convolutional layer which was computationally expensive as said before. So to reduce the computational cost the 5×5

convolutional layer, i.e. time-consuming and requires high computational power was replaced by two 3×3 convolutional layers as shown in Figure 5.3.1.5.

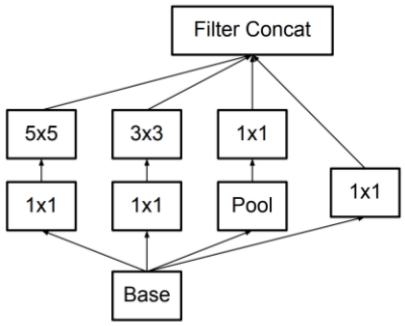


Figure 5.3.1.4: Inception v1 model

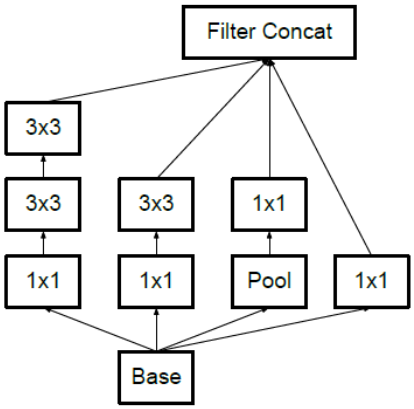


Figure 5.3.1.5: Inception v3 model

Replacing bigger convolutions with smaller convolutions definitely leads to faster training. Say a 5 × 5 filter has 25 parameters; two 3 × 3 filters replacing a 5 × 5 convolution has only 18 (3*3 + 3*3) parameters instead. By reducing the number of parameters involved in a network, the computational efficiency could be reduced too. It also keeps a check on the network efficiency.

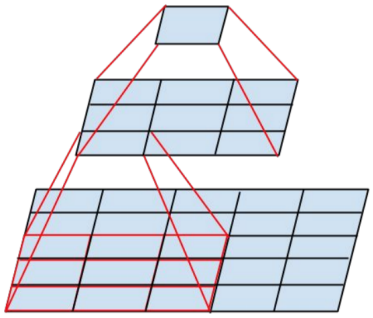


Figure 5.3.6: Diagram proving that the 3x3 convolutions reduce the number of parameters

2. Asymmetric convolutions

A 3 × 3 convolution could be replaced by a 1 × 3 convolution followed by a 3 × 1 convolution. If a 3 × 3 convolution is replaced by a 2 × 2 convolution, the number of parameters would be slightly higher than the asymmetric convolution proposed. The two-layer solution is 33% cheaper for the same number of output filters if the number of input and output filters is equal.

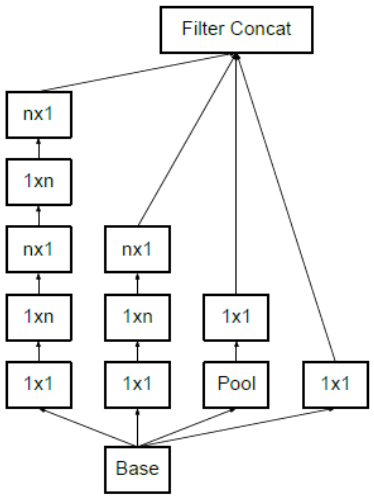


Figure 5.3.1.7: Structure of asymmetric convolutions

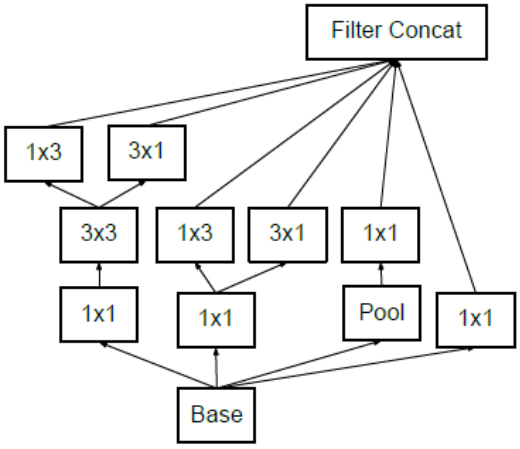


Figure 5.3.1.8: Asymmetric convolution of inception v3 model

3. Auxiliary classifier

An auxiliary classifier is a small CNN inserted between layers during training, and the loss incurred is added to the main network loss. It aims to improve the convergence of very deep neural networks and to combat the vanishing gradient problem in very deep networks. With auxiliary classifier, the network has higher accuracy, hence it acts as a regularizer in inception v3 model architecture.

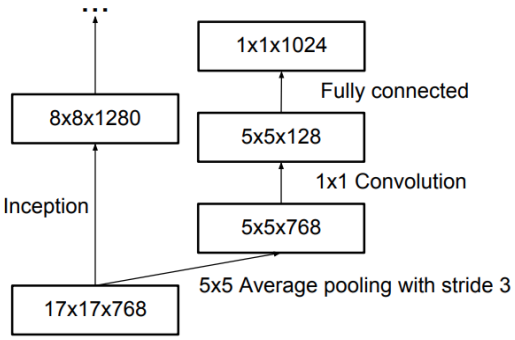


Figure 5.3.1.9: Auxiliary classifier

4. Grid size reduction

Grid size reduction is usually done by pooling operations. However, to combat the bottlenecks of computational cost, a more efficient technique is proposed:

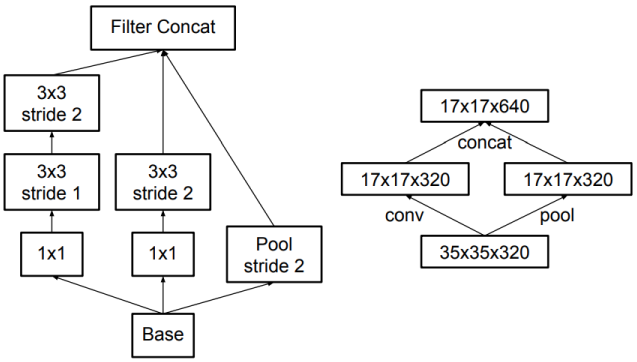


Figure 5.3.1.10: Grid Size Reduction

In summary, the model itself is made up of symmetric and asymmetric building blocks, including convolutions, average pooling, max pooling, concatenations, dropouts, and fully connected layers. Batch normalization is used extensively throughout the model and applied to activation inputs. Loss is computed using Softmax. All the above concepts are consolidated into the final architecture.

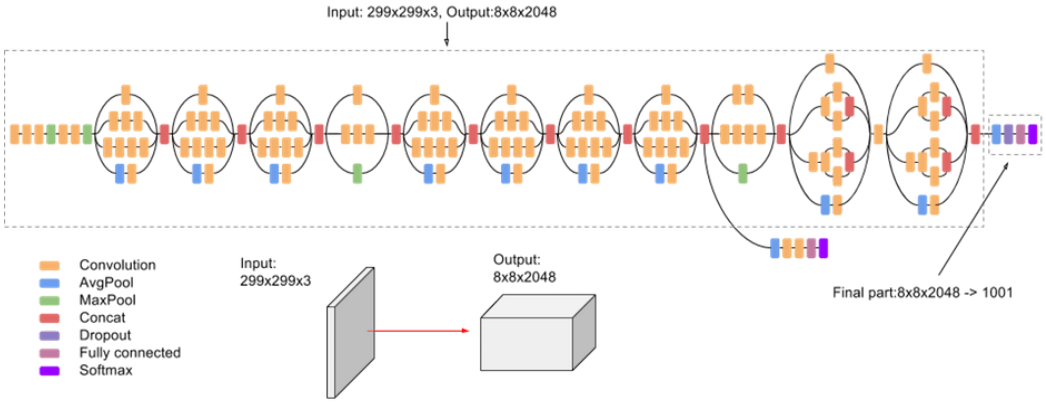


Figure 5.3.1.11: Inception v3 model

Machine Learning Models

Once the computation is done, you can observe the enhanced data in a Data Table. With the retrieved embeddings, we continue with the machine learning methods Orange offers. Machine learning models such as Logistic Regression, Tree, Random Forest, Gradient Boosting as well as Neural Network are built using embedded data from train dataset. The details and parameters used for each model are shown in Figure 5.3.1.12. As such we are able to train machine learning models without writing code.

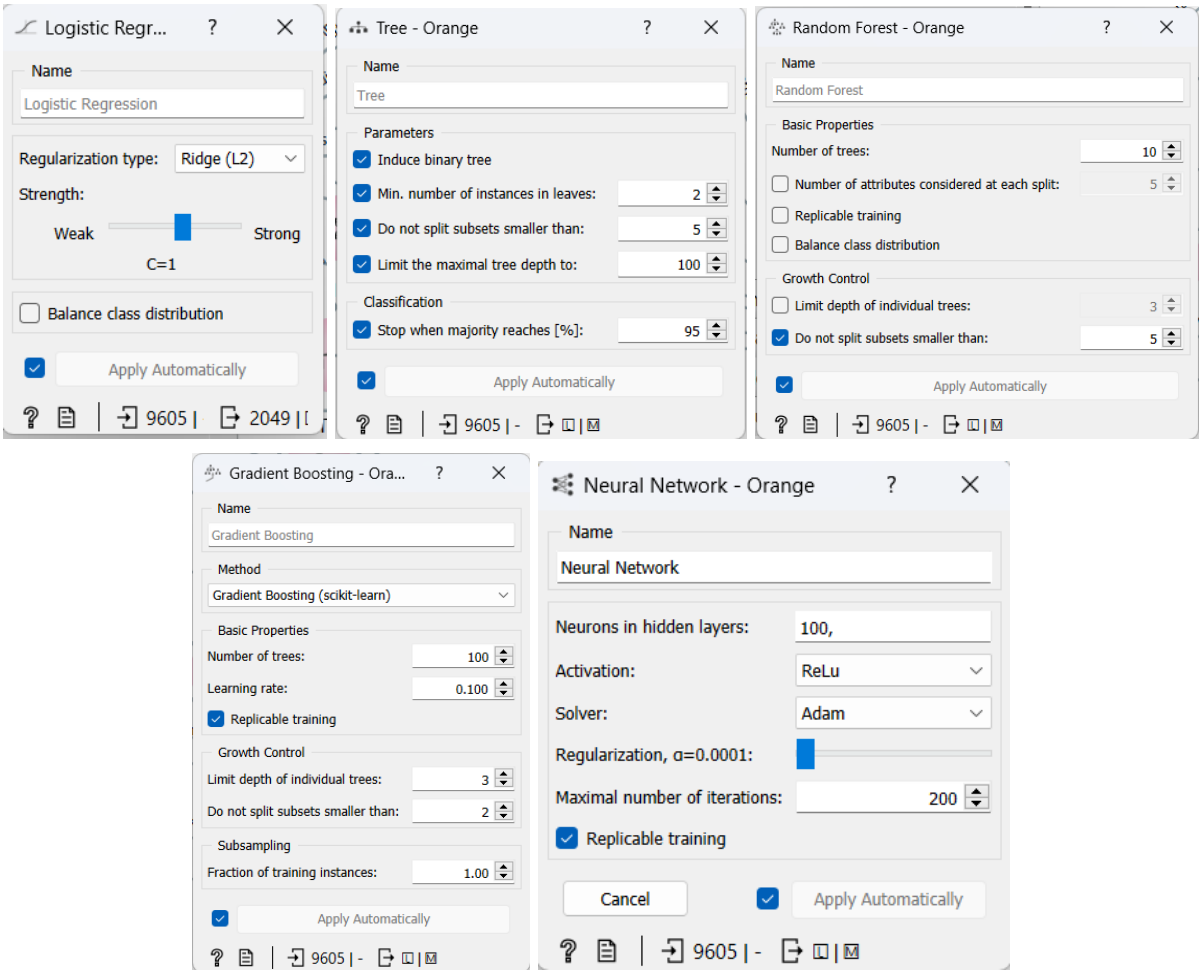


Figure 5.3.1.12: Machine learning models

Multiple models can be trained at the same time and we can easily compare evaluation results in a the table inside the widget ‘Test and Score’. The results for our models are as follows:

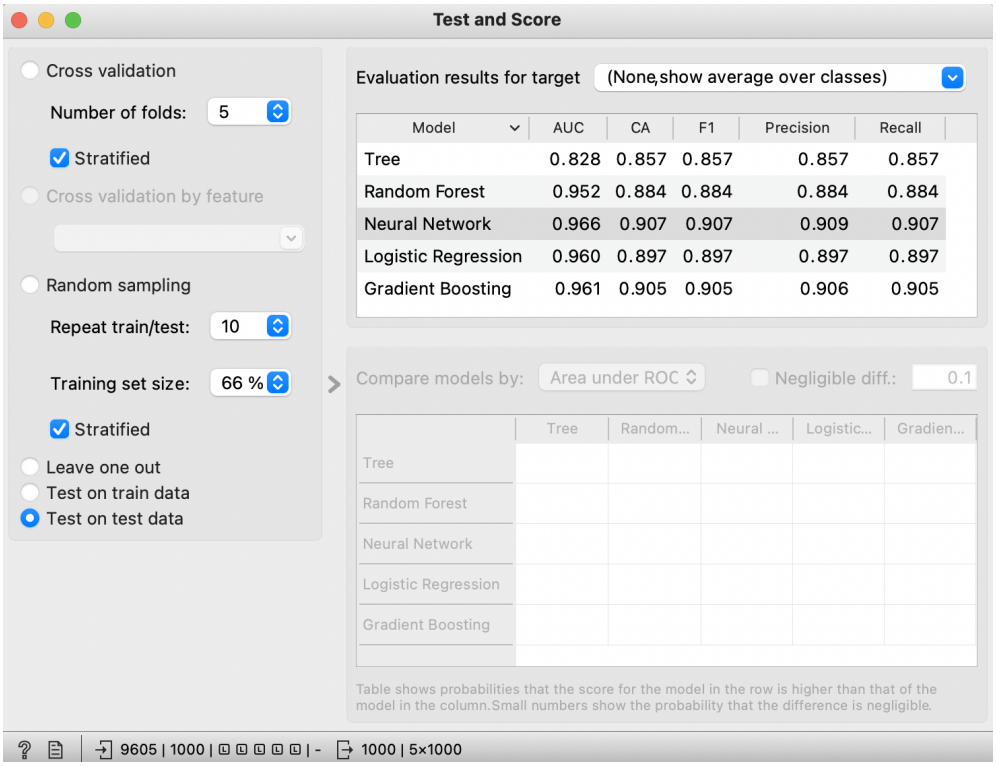


Figure 5.3.1.13: Test and score of models

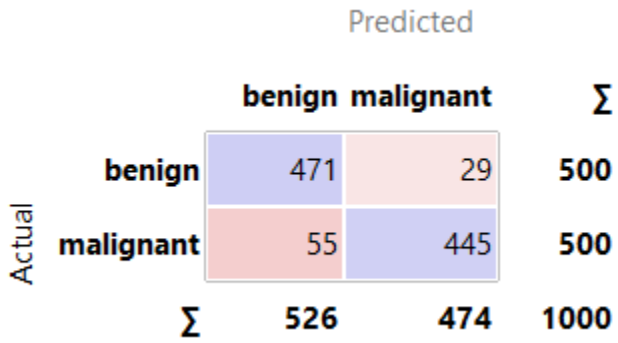


Figure 5.3.1.14: Confusion Matrix for NN Model

	Accuracy	False Negative Rate	Precision	Recall
Result	0.907	0.110	0.909	0.907

Figure 5.3.1.15: Summary of Results for NN Model

As observed, neural network performs the best as it has the highest accuracy of 0.907, lowest FNR of 0.110, highest precision of 0.909 and highest recall of 0.907. The results obtained are optimal and hence InceptionV3 is possibly another model that can be considered to be use for our application.

5.3.2 ResNet50

ResNet50 is a deep learning Residual Network model with 50 layers, whose weights were pre-trained on ImageNet.

Usually, the general neural network accuracy tends to decline as more layers are involved due to overfitting, occupied with vanishing gradients, in which the speed of learning decreases very rapidly for the shallower layers as the network trains. ResNet50 solved the problem with the skip connections, which perform identity mapping by adding the input of nth layer to the (n+1)th layer as below.

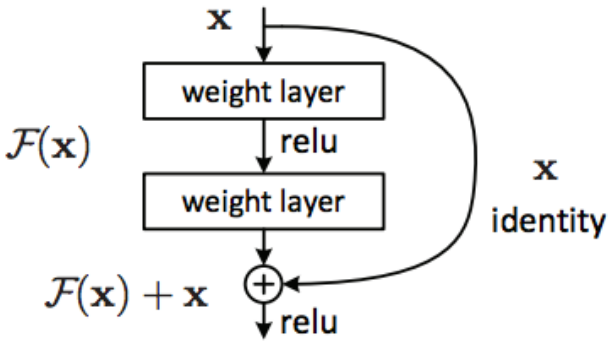


Figure 5.3.2.1.: Identify Mapping

In ResNet50, there are two main types of blocks, depending on whether the input/output dimensions are the same or different: the "identity block" and the "convolutional block" (Kaushik, 2020). The identity block is implemented when the input activation of the nth layer has the same dimensions as the output activation of the (n+2)th layer. In addition, batch normalizations are included to accelerate the training.

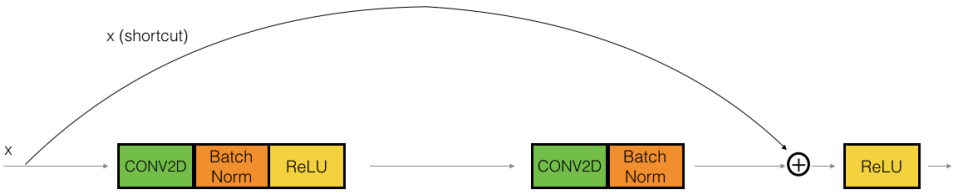


Figure 5.3.2.2: Batch Normalization

On the other hand, if the input and output dimensions are not identical, a convolution layer, for example, 1*1 convolution is added to the shortcut path as below to resize the input.

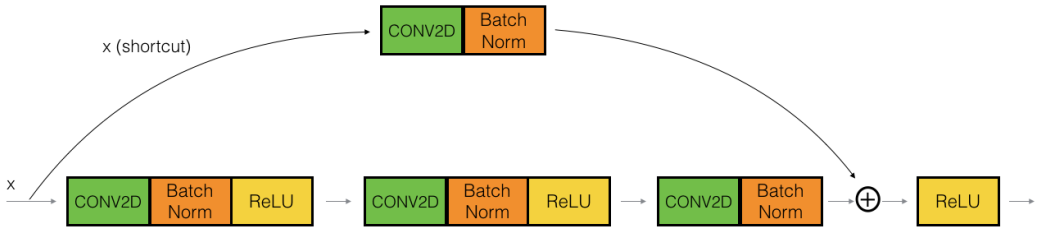


Figure 5.3.2.3: Shortcut Path

Specifically, ResNet-50 Embedding model to be implemented for melanoma prediction includes:

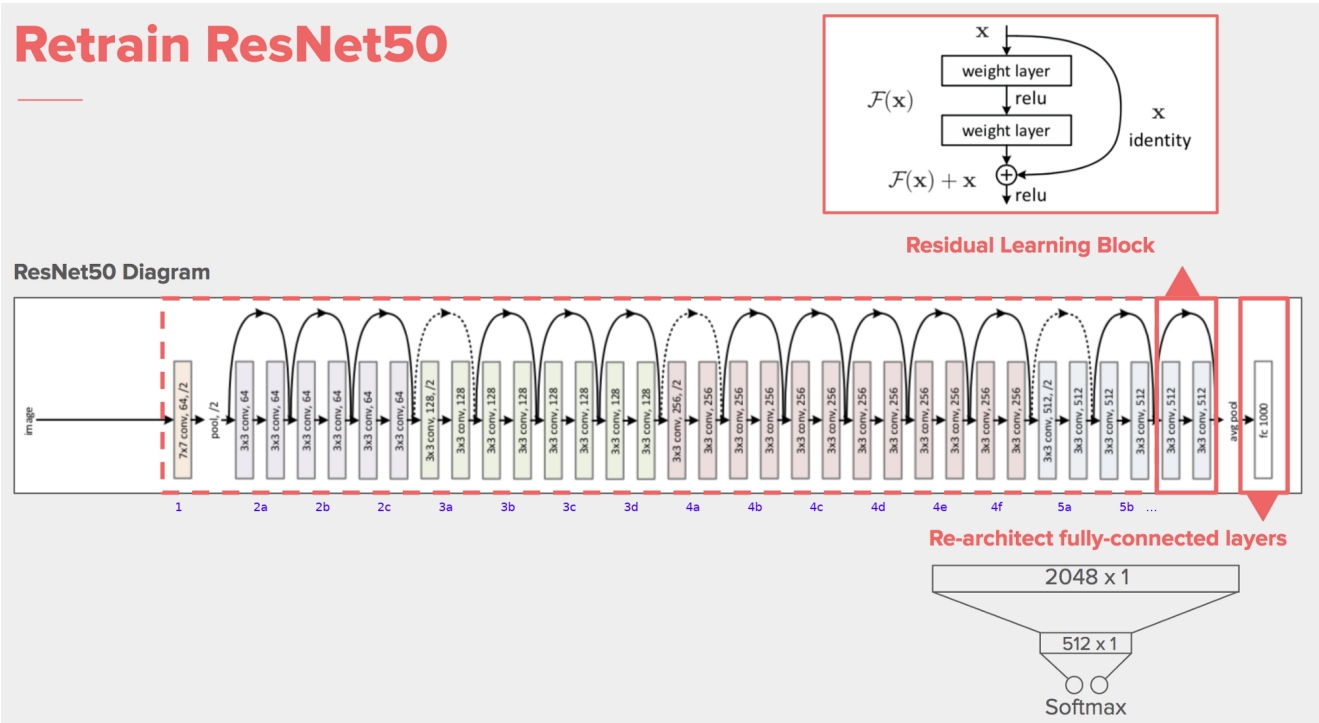


Figure 5.3.2.4: RestNet50 Architecture

- Zero-padding of size 3*3 to pad the inputs.
- Stage 1:
    - The 2D Convolution has 64 filters of shape (7,7) and uses a stride of 2.
    - BatchNorm is applied to the 'channels' axis of the input.
    - ReLU activation is applied.
    - MaxPooling uses a kernel size of (3,3) and a stride of 2.
- Stage 2 (2a, 2b, 2c): 1 convolutional layer + 2 identity layers
    - The convolutional block uses three sets of filters of size [64,64,256] with a stride of 1.

    - The 2 identity blocks use three sets of filters of size [64,64,256].
- Stage 3 (3a, 3b, 3c, 3d):
    - The convolutional block uses three sets of filters of size [128,128,512] with a stride of 2.
    - The 3 identity blocks use three sets of filters of size [128,128,512].
- Stage 4 (4a, 4b, 4c, 4d, 4e, 4f):
    - The convolutional block uses three sets of filters of size [256, 256, 1024] with a stride of 2.
    - The 5 identity blocks use three sets of filters of size [256, 256, 1024].
- Stage 5 (5a, 5b):
    - The convolutional block uses three sets of filters of size [512, 512, 2048] with a stride of 2.
    - The 2 identity blocks use three sets of filters of size [512, 512, 2048].
- The 2D Average Pooling uses a window of shape (2,2).
- The 'flatten' and normalized layer, followed by a fully connected layer to connect the embedded input to a simple model (e.g. logistic regression) for output.

Both the train and test dataset will be embedded through the ResNet-50 pipeline as above to be inputted in a logistic regression, which produced a confusion matrix as follows.

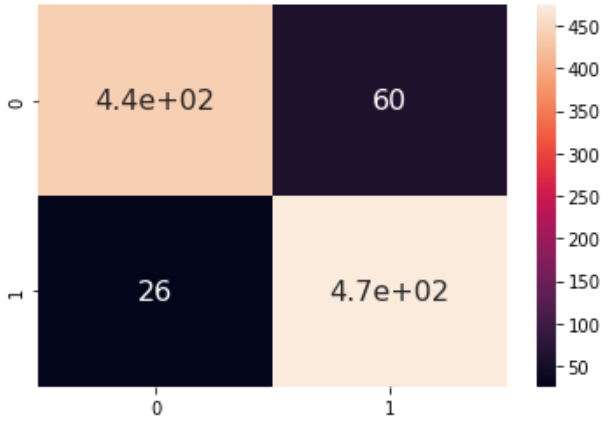


Figure 5.3.2.5: Confusion Matrix

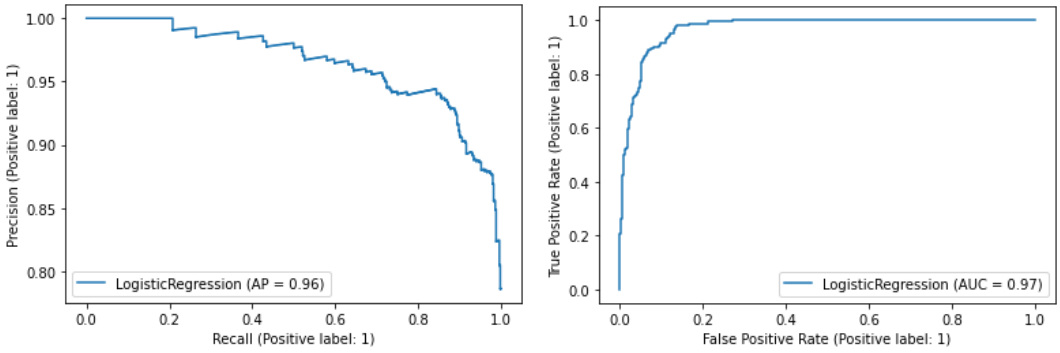


Figure 5.3.2.6: Precision/ Recall Curve & AUC Curve

	Accuracy	False Negative Rate	Precision	Recall
Result	0.914	0.052	0.967	0.907

Figure 5.3.2.6: Summary of Results for NN Model

Overall, ResNet-50 Embedding resulted in a relatively good performance, with 91.4% accuracy, recall of 94.8%, precision of 88.76%. In addition, the ROC as well as precision-recall curve also indicate a good tradeoff with the area under the ROC curve values at 0.97. ResNet-50 currently also has one of the lowest false negative rates of only 5.2%. This is extremely important as this suggests that the model predicts benign wrongly for very little time. Hence making ResNet-50 one of the strongest contender to be use as the model for our application

5.3.3 Deep AlexNet-Based CNN (PyTorch)

AlexNet is a popular CNN that was trained on subsets of ImageNet database used in the ILSVRC-2010 and ILSVRC-2012 competitions, winning top-5 test error rate of 15.3%, compared to 26.2% achieved by the second-best entry (Alex, 2012). The ImageNet database has over 15 million labeled, high-resolution images belonging to 22,000 categories. AlexNet is 8 layers deep and can classify images into 1000 categories, such as keyboard, mouse, pencil, etc.

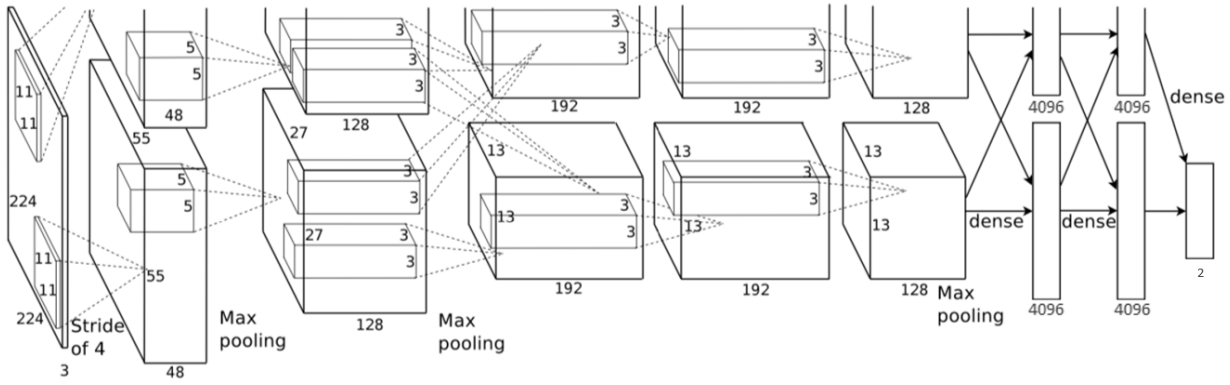


Figure 5.3.3.1: Illustration of AlexNet's architecture.

In this model, we will create a Convolutional Neural Network (CNN) inspired by the AlexNet model. The architecture consists of eight layers: five convolutional layers and three fully-connected layers. It uses the Activation function to increase nonlinearity and boost convergence rate and several GPUs for quicker training. The varied kernel sizes are one of the primary reasons the AlexNet could perform well on the image dataset. Such kernel size enabled the model to fully understand and record the essential features of the input data. In the proposed model, we have implemented similar kernel size variations in the different convolutional layers. These are some of the features used that are new approaches to convolutional neural networks:

1. **ReLU Nonlinearity** - AlexNet uses Rectified Linear Units (ReLU) instead of the tanh function, which was standard at the time. ReLU's advantage is in training time; a CNN using ReLU was able to reach a 25% error on the CIFAR-10 dataset six times faster than a CNN using tanh.

2. **Multiple GPUs** - Back in the day, GPUs were still rolling around with 3 gigabytes of memory (nowadays those kinds of memory would be rookie numbers). This was especially bad because the training set had 1.2 million images. AlexNet allows for multi-GPU training by putting half of the model's neurons on one GPU and the other half on another GPU. Not only does this mean that a bigger model can be trained, but it also cuts down on the training time.

3. **Overlapping Pooling** - CNNs traditionally "pool" outputs of neighboring groups of neurons with no overlapping. With overlapping, there is a reduction in error and models with overlapping pooling generally find it harder to overfit.

4. **Dropout Layers** - AlexNet also addresses the over-fitting problem by using drop-out layers where a connection is dropped during training with a probability of p=0.5. Although this avoids the network from over-fitting by helping it escape from bad local minima, the number of iterations required for convergence is doubled too.

```
AlexNet(
  (features): Sequential(
    (0): Conv2d(3, 96, kernel_size=(11, 11), stride=(4, 4), padding=(2, 2))
    (1): ReLU(inplace=True)
    (2): MaxPool2d(kernel_size=3, stride=2, padding=0, dilation=1, ceil_mode=False)
    (3): Conv2d(96, 256, kernel_size=(5, 5), stride=(1, 1), padding=(2, 2))
    (4): ReLU(inplace=True)
    (5): MaxPool2d(kernel_size=3, stride=2, padding=0, dilation=1, ceil_mode=False)
    (6): Conv2d(256, 512, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
    (7): ReLU(inplace=True)
    (8): MaxPool2d(kernel_size=3, stride=2, padding=0, dilation=1, ceil_mode=False)
    (9): Conv2d(512, 512, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
    (10): ReLU(inplace=True)
    (11): Conv2d(512, 256, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
    (12): ReLU(inplace=True)
    (13): MaxPool2d(kernel_size=3, stride=2, padding=0, dilation=1, ceil_mode=False)
    (14): AdaptiveAvgPool2d(output_size=(6, 6))
  )
  (classifier): Sequential(
    (0): Linear(in_features=9216, out_features=4096, bias=True)
    (1): Dropout(p=0.5, inplace=False)
    (2): Linear(in_features=4096, out_features=4096, bias=True)
    (3): Dropout(p=0.5, inplace=False)
    (4): Linear(in_features=4096, out_features=2, bias=True)
  )
)
```

Figure 5.3.3.2: Model Summary of our AlexNet-based architecture.

From the model summary, the first stage consists of the 2D Convolution which has 3 channels in the input image with 96 channels produced by the convolution. Kernel Size is (11,11). Input Shape is (227, 227) and uses a stride of (4, 4), padding of (2, 2). ReLU activation is applied.

MaxPooling uses a kernel size of (3,3) and a stride of 2. As we move along the stages, note that the kernel size constantly changes as explained above. AdaptiveAvgPool2d applies a 2D adaptive average pooling over an input signal composed of several input planes and produces a targeted output size of (6, 6). For training, we will be using 10 epochs and a relatively smaller batch size of 64.

After defining our model, we want to avoid overfitting problems while training the model as it may cause the accuracy when testing on our test set to be much lower than that on the train set. Hence, we carry out a K-fold cross-validation with K=5 to detect overfitting. We split the data points into 5 equally sized subsets in K-folds cross-validation, called "folds." One split subsets act as the testing set, and the remaining folds will train the model. After partitioning the data into 5 folds, we iteratively train the algorithm on 4 folds while using the remaining holdout fold as the test set. This method allows us to tune the hyperparameters of the neural network or machine learning model and test it using completely unseen data.

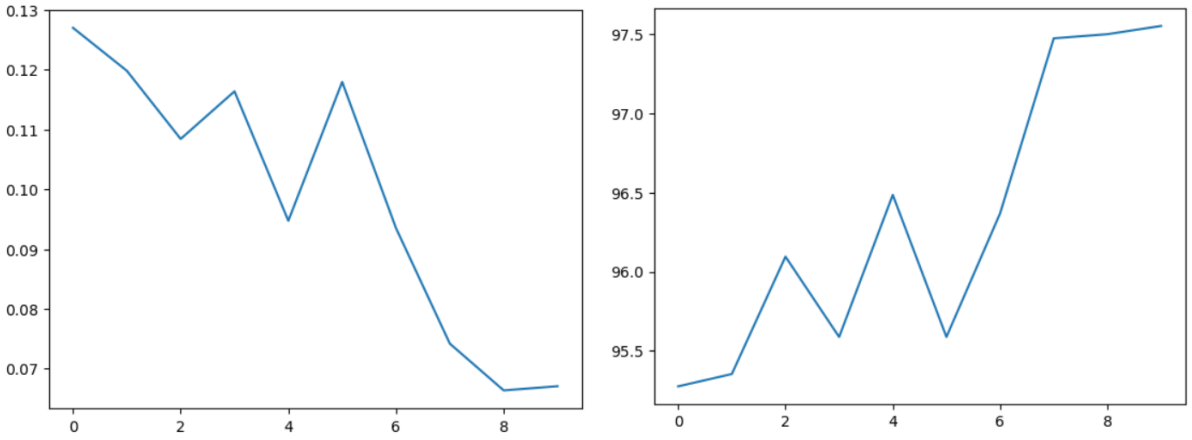


Figure 5.3.3.3: Loss Graph (Left) & Accuracy Graph (Right)

From the graphs above, we can see that for 10 epochs and 5 K-fold validation, we were able to achieve a downwards sloping loss graph and an upward sloping accuracy graph. However, as we had only train the model on 10 epochs, we were unable to obtain a tail to our graphs which thus suggest that the model could still be underfitted. Thus further improvements could be achieved with the model if higher epochs were used during training.

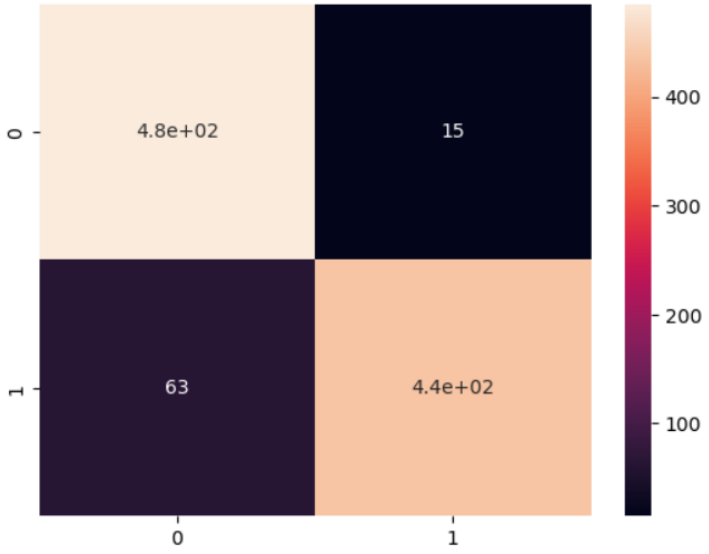


Figure 5.3.3.4: Confusion Matrix

	Accuracy	False Negative Rate	Precision	Recall
Result	0.922	0.126	0.967	0.874

Figure 5.3.3.5: Summary of Results

With the trained AlexNet based model, we test the model using the test dataset and produce a confusion matrix as seen above. From the confusion matrix, calculations indicate that the AlexNet based model produced an outstanding performance, with 92.2% accuracy, recall of 87.4%, precision of 96.7%, false negative rate of 12.6%. From the results above, we observe that AlexNet currently has the highest accuracy and highest precision. Thus, it is highly successful in the general classification of melanoma thus also making it one of the stronger contender to be use as the model for our application.

6. Analysis & Evaluation

6.1 Comparison

ResNet and Orange are trained on several machine learning models, including logistic regression, tree and random forest. AlexNet-Based and TensorFlow are trained on neural network models.

Accuracy are compared within ResNet and Orange and we found out that neural network has the highest accuracy. Hence we move on to compare neural networks between each architecture model, namely, TensorFlow Default Model, TensorFlow Improved Model, Orange, ResNet5 and AlexNet-Based.

6.2 Significance of performance metrics

With the correct metrics, they can effectively increase the models' predictive power. Before diving into each metrics, data dictionary of important terms are explained below:

True Positive (TP)	Predicted malignant when the actual diagnosis is malignant
True Negative (TN)	Predicted benign when the actual diagnosis is benign
False Positive (FP)	Predicted malignant when the actual diagnosis is in fact benign
False Negative (FN)	Predicted benign when the actual diagnosis is in fact malignant

The following are the several metrics used to measure the performance of models:

Accuracy

The accuracy of models shows the models' ability to obtain correct predictions (a true positive and a true negative).

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + False\ Negative + False\ Positive + True\ Negative}$$

It is important for the machine learning model to predict the severity of melanoma. High accuracy means better decisions can be made overall. Correctly identify the patients who suffer from benign would allow them to have early treatment before the disease worsens. For malignant, treatment could be done immediately to increase the survival rate of patients.

Inaccuracy can lead to incorrect assumptions, this can mean missing opportunities for early, potentially life-saving, treatment. This would put a toll on health care and patient costs. Hence the higher the accuracy, the higher the models' predictive power. Additionally, accuracy is the metrics use to evaluate the effectiveness of classification problems, thus it is one of the more important metrics used in the evaluation of which model is the best.

FNR

FNR is the rate for which the model falsely predicted the positive class (malignant) labels to be negative.

$$FNR = \frac{False\ Negative}{True\ Positive + False\ Negative}$$

A false negative melanoma test indicates that a person is predicted to have benign melanoma when the person actually suffers from malignant melanoma. Possible reasons why the machine learning might not recognise melanoma might be lack of irregular pigmentation and shape, altered sensation, the presence of inflammation and size < 7 mm.

Reduction of false negative rate is extremely important when it comes to life-threating disease, such as melanoma that has very low survival rate for stage III or IV melanoma. A finding from Melanoma Research Alliance showed that the 5-year survival rates for Stage III Melanoma and Stage IV Melanoma are 63.6% and 22.5% respectively. As such, if our model has a high false negative rate, this suggests that our model predicts many cases of benign when in actuality, the patient could be suffering from a maglinant melanoma which could be in Stage III or IV. This is extremely dangerous as we will be giving the patient a false sense of security and it prevents early treatment of melanoma. As a result of this wrong prediction, it could lead to patient not realising that his melanoma is cancerous and thus may not seek early treatment which could lead to the worsening of his condition. Therefore, in the healthcare sector, it is extremely important that we keep the false negative rate as low as possible and arguably, it is as important as a metric as compared to accuracy.

Precision

Precision is the ratio between the True Positives and all the Positives. We use precision to determine how correct the prediction of types of melanoma.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Precision is the measure of patients that we have correctly identified to have malignant melanoma out of all the patients that actually have malignant melanoma. It thus measures how ‘precise’ our model actually is at predicting positive cases. A high precision implies that our model is able to efficiently capture almost all the positive malignant cases. Therefore, if our priority is to ensure that all malignant cases are captured efficiently then precision is the metric that we should prioritise

A model with low precision means that it has a higher number of false positives. This is undesirable when treatment for malignant melanoma is being provided to patients with benign melanoma, resulting in the waste of medical resources. Hence the higher the precision, the more desirable it is.

Recall

Recall is the proportion of actual positives that was identified correctly. It is the measure of our model correctly identifying True Positives.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Recall is especially useful when we want our model to be highly susceptible to wrongly predicting benign cases. A low recall means that there is a high number of false negative. As mentioned in 2), it is highly undesirable and severe, to wrongly predict a patient to be not have cancer when in actuality he has cancer. Therefore, we want to have high recall to ensure that for all the patients that actually have malignant melanoma, they are not missed out in the prediction. Hence the higher the recall, the more desirable it is

Choice of Metrics

Accuracy alone is insufficient to evaluate the models *(Phy, 2019)*, other metrics are required to make sure the model chosen is reliable and suitable for its context. It is extremely costly if we have false-negatives, where patients are predicted to suffer from benign melanoma when they actually suffer from malignant melanoma. This will delay early treatment for the patients who can potentially lead to full recovery. Hence, for the purpose of this project, the emphasis will be on accuracy, as well as a low FNR.

As a result, more focus will be place on recall, rather than precision since false negative is more costly than false positive. “Anything that doesn’t account for false negative is a crime” in the healthcare industry and this is because, the cost of not treating a patient that has cancer is extremely costly, the cost is potentially a lost of a precious life. Therefore, among these metrics, more emphasis will be put on **FNR, Recall and Accuracy,** when evaluating which one is the best performing model.

6.3 Choice of Model

The following are the summary of results for all five models:

	Accuracy	FNR	Precision	Recall
TensorFlow Default Model	0.694	0.610	0.990	0.390
TensorFlow Improved Model	0.913	0.114	0.937	0.886
Orange	0.907	0.093	0.909	0.907
ResNet50	0.914	0.052	0.890	0.948
AlexNet-Based	0.922	0.126	0.967	0.874

Table 6.3.1: Summary of Results

From the table above, we can observe that AlexNet-Based has the highest accuracy with 0.922 and the highest precision at 0.967. ResNet50, on the other hand, has the lowest FNR of 0.052

and highest recall of 0.948. Therefore, our choice of model is either between ResNet50 or AlexNet.

After evaluating, our group has decided to use **ResNet50 as our chosen model** for our application. As previously mentioned, in the context of healthcare, it is extremely important to have low false negative rates and our chosen model should as much as possible be able to detect almost all positive cases of Melanoma so that no patients that have a cancerous melanoma is left undetected and untreated. Therefore, a low FNR and high recall is preferred, in which case ResNet50 is the most optimal choice. Additionally, if we compare the accuracy of ResNet50 to AlexNet, we see that there is only a 0.008 point difference which is extremely negligible. Moreover, the AlexNet model that was trained is actually possibly an underfitted model as the model was trained with only 10 epochs with 5 cross validation, thus, compared to an established CNN model like ResNet50, it may also be more advisable to also use ResNet.

Additionally, according to an article posted on the Guardian, a small study published in 2018 found that dermatologists accurately diagnosed Melanoma (sensitivity) with **86.6% and 88.9%** accuracy, depending on the stage of the Melanoma. Just over half the dermatologists were at “expert” level with more than five years of experience, 19% had between two and five years’ experience, and 29% were beginners with less than two years under their belt (Agence, 2018). Therefore, it can be seen that with our ResNet model that was trained on close to 10,000 images, our model is already able to outperform these dermatologists that had a wealth of experience. Both in terms of accuracy and sensitivity (recall), all our model (except the default TensorFlow Model) already achieve higher results. Therefore, we are confident that our choice to use ResNet50 is ideal and is comparable to current industry standards.

Moreover, according to the article, it was found that the dermatologists’ performance improved when they were given more information of the patients and their skin lesions. The team said AI may be a useful tool for faster, easier diagnosis of skin cancer, allowing surgical removal before it spreads. Therefore, with the use of our ResNet50 model, it will potentially be able to aid dermatologists in the faster, more accurate diagnosis of melanoma.

7. Business Implementation

7.1 Novelty of Application

The novelty of our idea is based on the fact that Melanoma is easy to self-diagnose and exists on the outer skin (Halpern, 2020) which allows patients to take pictures from the comfort of their smartphones. Furthermore, our application significantly reduces the cost of Melanoma diagnosis, which improves the ease of access to Melanoma diagnosis and encourages early treatment. Early detection and recognition of skin cancer are very important (Halpern, 2020). Recognizing the early warning signs of melanoma and doing regular self-examinations of your skin can determine the severity of melanoma early, when the disease is more curable. Through our application and machine learning model, patients will know about their existing condition early and aid decision making in the treatment process, whether it is to perform self-treatment or get professional help from doctors.

7.2 Introduction of Application

For this project, we have developed **TrueHealth**, a machine learning web application hosted on Heroku that detects and diagnoses the severity of Melanoma. Our application consists of our **ResNet50** model, hosted on our Flask Backend to provide diagnosis predictions based on user-uploaded images. Our application is designed to be simplistic, intuitive and user-friendly, such that non-professionals like patients are capable of using the application for the purpose of self-diagnosis (White, 2022). Furthermore, dermatologists can use our application to get a second opinion or expedite the diagnosis process, thereby increasing the probability of a correct diagnosis and the expedition of the treatment process.

TrueHealth is designed to be used online that is both web-friendly and mobile-friendly. In 2022, our targeted market country of Australia is a developed country that currently has a smartphone penetration rate of 86% (Statista, 2022) and internet penetration rate of 91% (Statisa, 2022). These rates are predicted to increase over the next few years (Gjorgievska, 2022), providing more support for our intended usage scenarios. Leveraging on this nation wide access to technology, TrueHealth rides on this wave of increasing trend of reliance on technology to implement our self-diagnosis application.

7.3 Value-added Features

From our research on the existing technology to diagnose Melanoma, the medical tools available in hospital are not accurate or fast enough (Kadumpur, 2019). This leads to a poor and late diagnosis that inherently affects the treatment process, which is crucial as Melanoma has to be treated as early as possible to prevent the exponential spread of the cancer cells. With this in mind, TrueHealth hinges on the value-added features of early, fast detection and severity diagnosis to aid decision making.

It is found that a skilled dermatologist usually follows a series of steps, starting with naked eye observation of suspected lesions, then dermoscopy (magnifying lesions microscopically) and

followed by biopsy. This would consume time and the patient may advance to later stages. Moreover accurate diagnosis is subjective, depending on the skill of the clinician. It is found that the best dermatologist has an accuracy of less than 80% in correctly diagnosing melanoma. Adding to these difficulties, there are not many skilled dermatologists available globally in public healthcare. (Kadumpur, 2019)

7.4 Technology Stack

TrueHealth is developed with React.js for the frontend design, with Node.js for the backend and a Flask server that holds our machine learning model. Our technology stack architecture is as shown below.

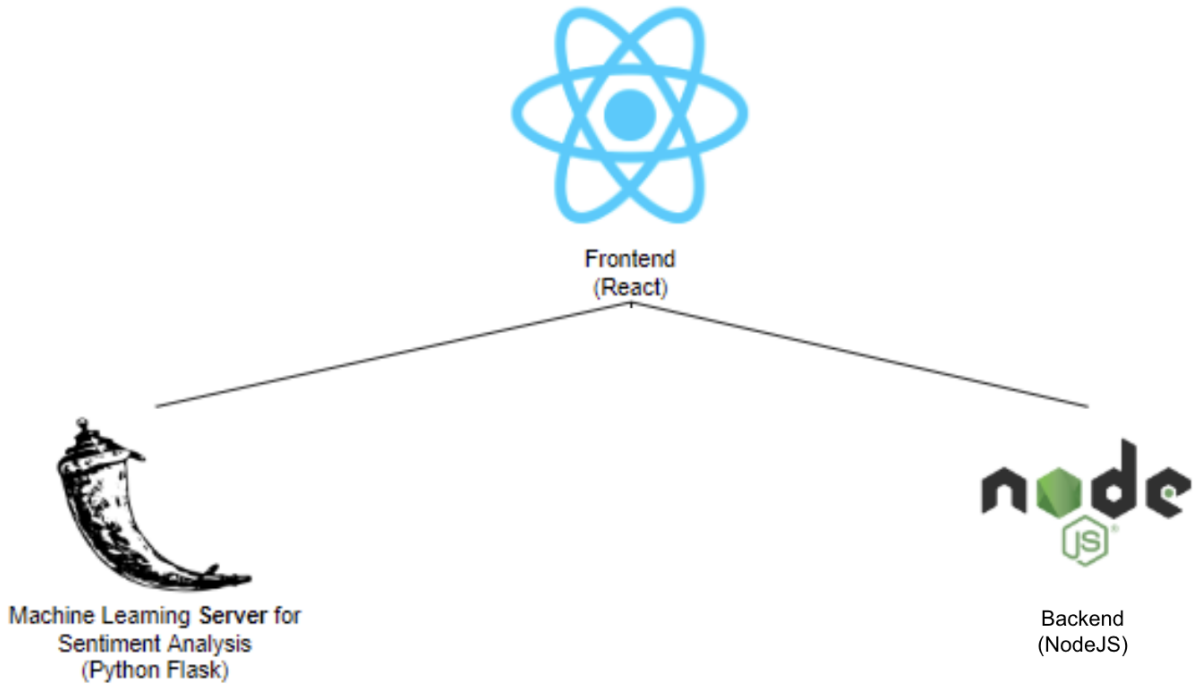


Figure 7.4.1: Technology Stack Architecture

7.5 Features of Application

TrueHealth consists of a 3-stage process of taking photo, diagnosing and analysing, serving the extensive use cases of both patients-at-home and professional dermatologists. The below shows the user journey of our application.

7.5.1 Uploading potential Melanoma image using camera

With the camera of a smartphone or web-camera of a computer, the user can simply take an close-up snapshot of their area of concern. After uploading the JPEG image, the Melanoma

image is displayed on our medical image viewer. Our viewer shows information like the zoom level (bottom right) and the brightness level (bottom left).

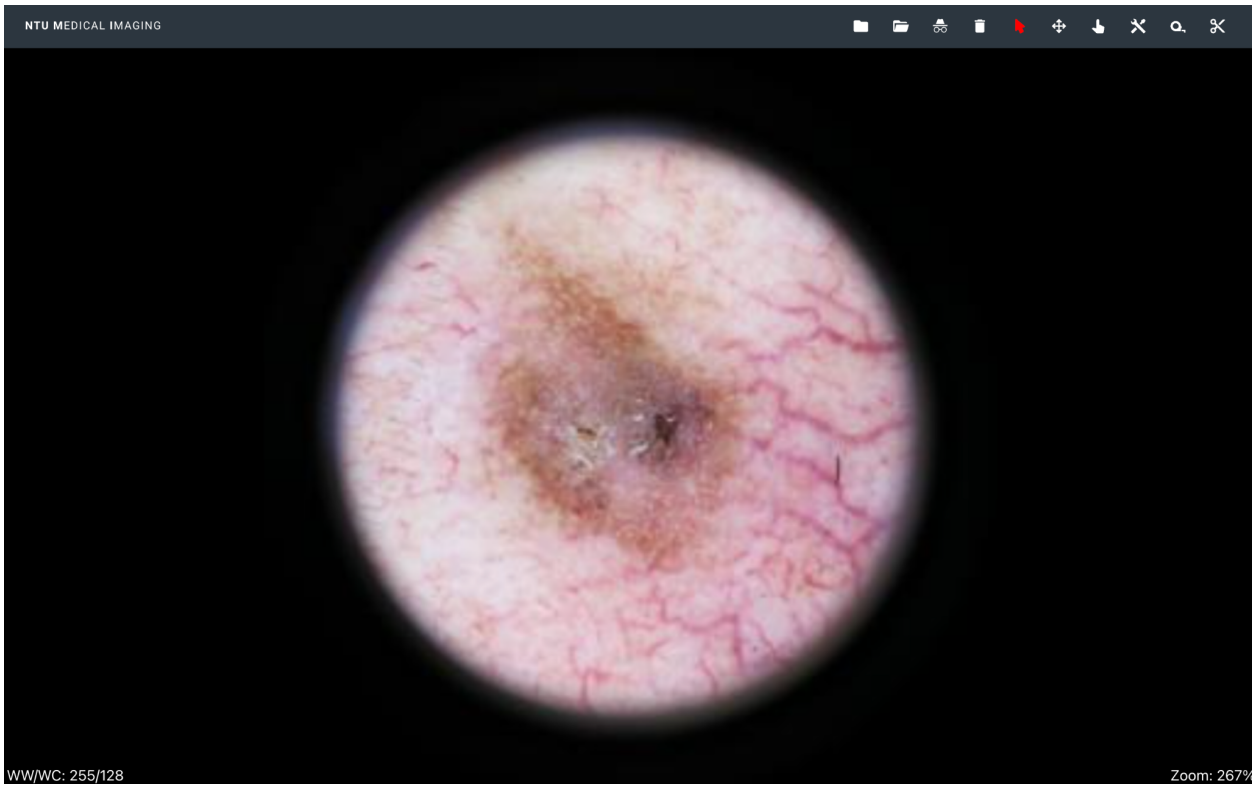


Figure 7.5.1.1: Image uploaded through application

7.5.2 Diagnosing Melanoma using our ResNet50 model

With the image uploaded, the user selects the “Diagnose Melanoma” button. This sends the image to our backend server hosted on Flask. Our backend server contains our trained ResNet50 model that takes in the uploaded image as input and outputs a prediction of whether the uploaded case of Melanoma is benign or malignant. This prediction is returned to our frontend and shown to the user.

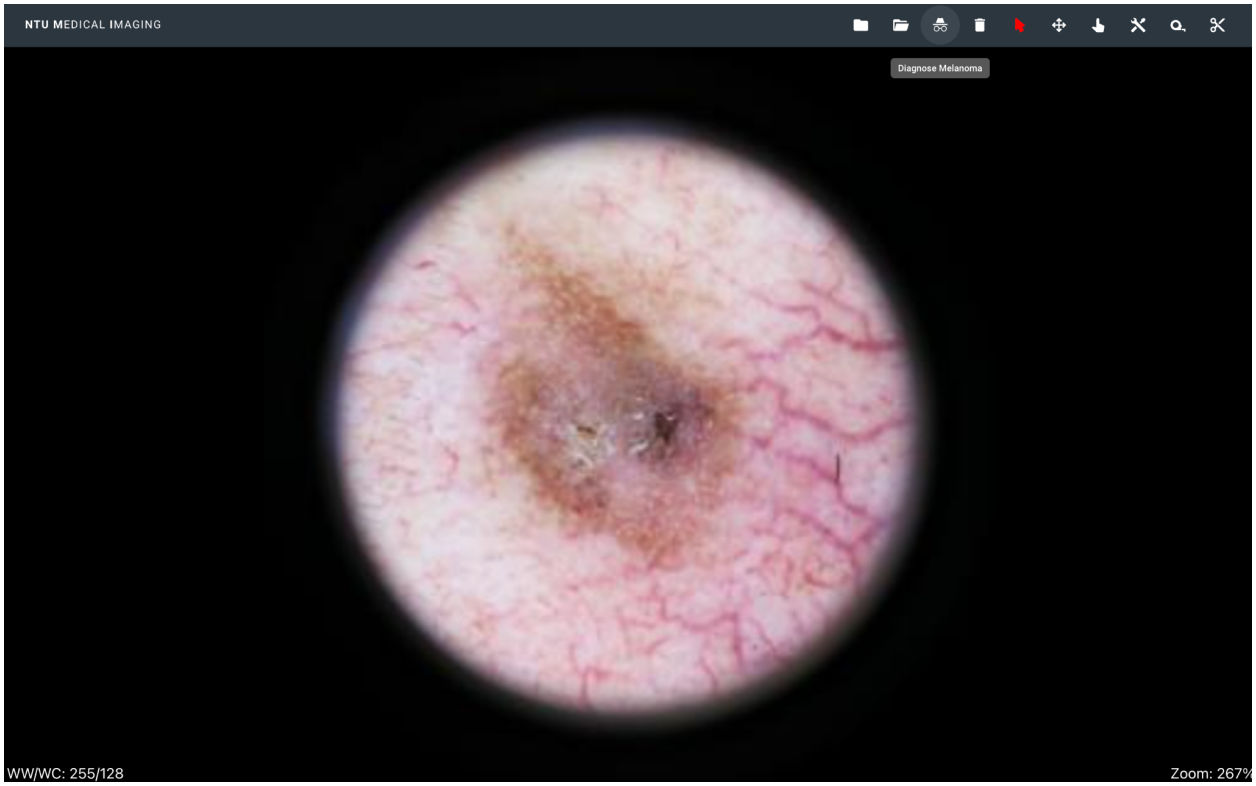


Figure 7.5.2.1: Diagnose melanoma tool in application

7.5.3 Post-Diagnosis Analysis

With the diagnosis in hand, the user can proceed to make further analysis using the array of tools integrated into our application. These tools are imported from Cornerstone library, which hosts multiple medical analysis tools. The tools included are general tools such as WW/WC, zooming and panning, measurement tools with annotation field and measurement tools for length, area, angle, elliptical, rectangle and Freehand ROI. With the user in mind, we also have a ‘Clear All’ feature to reset the application and allow further uploads.

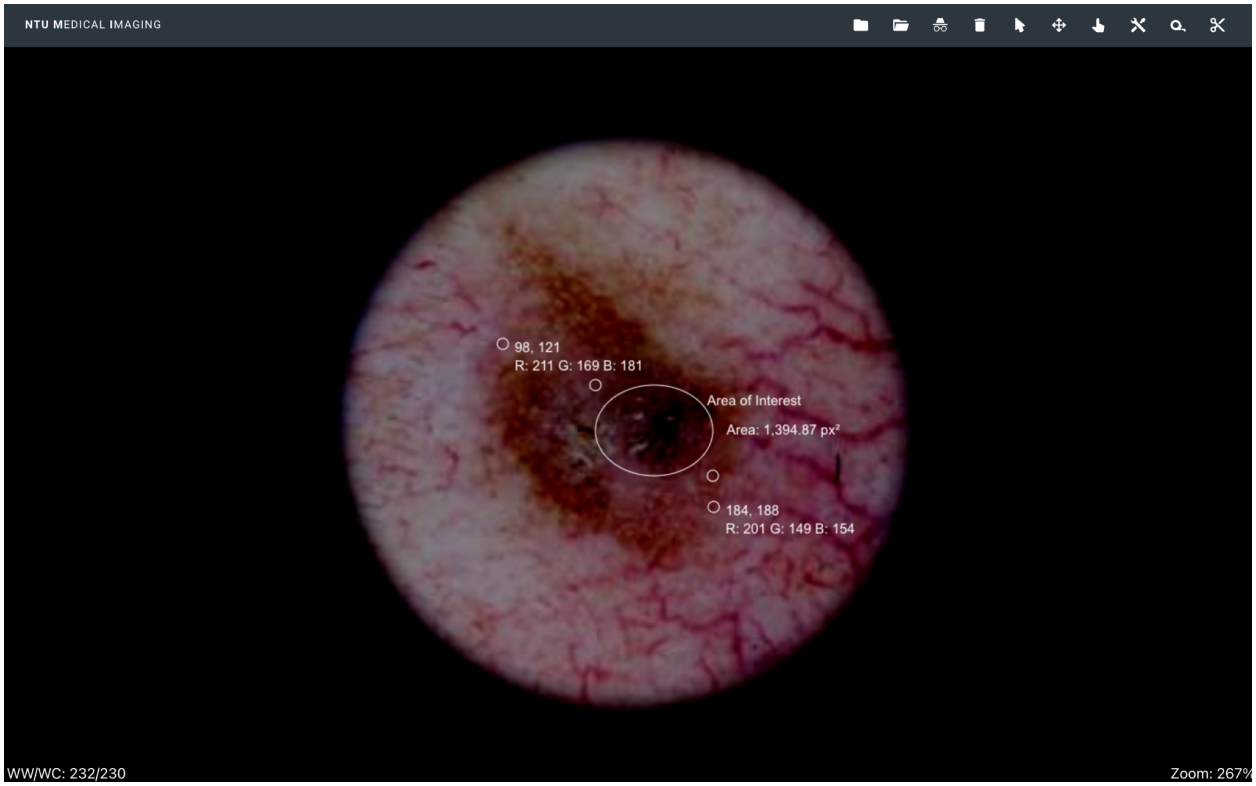


Figure 7.5.3.1: Additional tools in application

7.6 Business Model

As our application's target market includes users who are unable to afford proper medical diagnosis of Melanoma, our services will be provided free-of-charge to increase its outreach and accessibility to users.

In order to ensure continuous development and sustainability of our application, our application will adopt the advertising business model, which is revenue generation strategy supported by the sale of advertising. In general, our application leverages on the audience that interacts with our application and revenue is earned by selling access to that audience.

8. Conclusion

8.1 Limitations

The following is the limitations of our application:

1. **Accuracy of our Diagnosis**

   Despite having a higher accuracy of diagnosis than medical institutes, there are more established tele-health competitors in the market that hold a higher accuracy than our application. Their machine learning model is trained on a wider range of data and they possess more specialised equipment like dedicated GPUs which help to increase the speed of training.

2. **Sustainability of our application**

   With more users using our application, there will be more costs incurred such as server and hosting costs for our application to be able to handle the increase in web traffic. Hence our business model of providing free access may not translate to sufficient revenue for us to maintain of our application. Furthermore, the advertising model requires an application with very high traffic and consistent high user count to attract advertisers.

3. **Limitations of Machine Learning Models**

   Even though deep learning has proven to show promising results in melanoma detection, they are still many limitations to it *(Engati, 2021)*:

   1) Abundant pure, unbalanced and large training datasets needed for CNN to be effective. This could be easily addressed by data augmentation on the training set of images.

   2) CNN has a hard time classifying images with different positions as it fails to encode the position and orientation of objects.

   3) Ugly Duckling recognition method is hardly being utilised in CNN. This recognition strategy is based on the concept that most normal moles on your body resemble one another, while melanomas stand out like ugly ducklings in comparison *(Skin Cancer Foundation, 2022)*.This highlights the importance of not just checking for irregularities, but also comparing any suspicious spot to surrounding moles in terms of size and colour, to determine whether it looks different from its neighbours.

   4) Speed of training a model is very slow when we have additional datasets to train the model using GPU with low computational capability. CNNs tend to be much slower because of operations like maxpool. In case the CNN is made up of

multiple layers, the training process could take a particularly long time if the computer does not have a good GPU.

8.2 Future plans

In the future, our application aims to expand to medical conditions beyond Melanoma. Our first stage of expansion will include other skin conditions such as precancerous lesions (blue and dysplastioc nevus, etc.), benign formations (moles, angeoma, dermatofibroma, etc.) and papilloma virus (warts, papillomas, mollusks, etc.). Our second stage of expansion will include diagnosis of medical conditions beyond the skin such as searching for traces of eye disorders such as Leukocoria or “white eye” through our similar use case of taking a picture. Our third stage of expansion will include diagnosis of medical conditions beyond the inputs of an image, such as a risk assessment of symptoms. For example, a user can simply enter his symptoms in text form into our risk assessment form, and our applications will be able to predict the possible causes for his symptoms, a summary of present and absent symptoms and a triage recommendation. Furthermore, our application will also incorporate computer vision technology to diagnose Musculoskeletal conditions. Our end goal is to develop a medical diagnosis application that is able to handle most medical conditions with optimal accuracy and speed, thus providing extensive use cases to our users.

8.3 Concluding Statement

In conclusion, our project utilised different types of convolutional neural networks (CNN) models. We also configured the parameters of the models appropriately to improve the accuracy and efficiency of each model. We found that the Resnet50 model performs the overall best in predicting whether a melanoma image is malignant or benign.

With this model, we are able to accurately and rapidly detect and determine the severity of Melanoma. As skin cancer is a dangerous and widespread disease, it demands early and fast detection. Without an easily accessible diagnosis solution in the market, our solution thus addresses the market gap by providing a free-to-use application that solves the issue of early diagnosis, with improved accuracy.

9. References

Admin. (2021, August 29). How to choose cross-entropy loss function in Keras? Knowledge Transfer. Retrieved November 11, 2022, from https://androidkt.com/choose-cross-entropy-loss-function-in-keras/

Advanced Guide to Inception v3 | Cloud TPU |. (n.d.). Google Cloud. https://cloud.google.com/tpu/docs/inception-v3-advanced

Agence, F. P. (2018, May 29). Computer learns to detect skin cancer more accurately than doctors. The Guardian. Retrieved November 12, 2022, from https://www.theguardian.com/society/2018/may/29/skin-cancer-computer-learns-to-detect-skin-cancer-more-accurately-than-a-doctor

Alex, K. (2012). ImageNet Classification with Deep Convolutional Neural Networks. Retrieved November 12, 2022, from https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

Awati, R. (2022) What are convolutional neural networks?, SearchEnterpriseAI. TechTarget. Available at: https://www.techtarget.com/searchenterpriseai/definition/convolutional-neural-network (Accessed: November 10, 2022).

Basta, N. (2020, April 5). The differences between sigmoid and Softmax activation function. Medium. Retrieved November 11, 2022, from https://medium.com/arteos-ai/the-differences-between-sigmoid-and-softmax-activation-function-12adee8cf322

Bioinformatics Laboratory, University of Ljubljana. (n.d.-a). Data Mining. https://orangedatamining.com/

Bioinformatics Laboratory, University of Ljubljana. (n.d.-b). Image Viewer. https://orangedatamining.com/widget-catalog/image-analytics/imageviewer/

Bioinformatics Laboratory, University of Ljubljana. (n.d.-b). Image Embedding. https://orangedatamining.com/widget-catalog/image-analytics/imageembedding/

Biswal, A. (2022) Convolutional Neural Network tutorial [update], Simplilearn.com. Simplilearn. Available at: https://www.simplilearn.com/tutorials/deep-learning-tutorial/convolutional-neural-network#:~:text=A%20convolutional%20neural%20network%20is,classify%20objects%20in%20an%20image (Accessed: November 10, 2022).

Blog, S. K. O. T. S. D. S. (2020, August 7). Video: Image embedding using deep learning with Python (DLPy) and SAS Viya. The SAS Data Science Blog. https://blogs.sas.com/content/subconsciousmusings/2020/08/07/dlpy-image-embedding/

Brownlee, J. (2019) How to configure image data augmentation in Keras, Machine Learning Mastery. Available at: https://machinelearningmastery.com/how-to-configure-image-data-augmentation-when-training-deep-learning-neural-networks/ (Accessed: November 10, 2022).

Convolutional Neural Network (2021) Engati. Available at: https://www.engati.com/glossary/convolutional-neural-network#:~:text=Some%20of%20the%20disadvantages%20of,classifying%20images%20with%20differnet%20positions (Accessed: November 10, 2022).

Gjorgievska, L. (2022a, November 1). 26 Sensational Statistics on the Most Popular Sports in Australia. Take a Tumble. https://takeatumble.com.au/insights/lifestyle/internet-usage-statistics/

Gjorgievska, L. (2022b, November 1). 26 Sensational Statistics on the Most Popular Sports in Australia. Take a Tumble. https://takeatumble.com.au/insights/lifestyle/internet-usage-statistics/

Gordon, L.G. et al. (2022) Estimated healthcare costs of melanoma and keratinocyte skin cancers in Australia and Aotearoa New Zealand in 2021, International journal of environmental research and public health. MDPI. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8948716/ (Accessed: November 12, 2022).

How accurate is dermoscopy compared to visual inspection of the skin for diagnosing skin cancer (melanoma) in adults? (n.d.). Cochrane. https://www.cochrane.org/CD011902/SKIN_how-accurate-dermoscopy-compared-visual-inspection-skin-diagnosing-skin-cancer-melanoma-adults

How to pick the optimal image size for training convolution neural network? (no date) https://www.raravind.com. Available at: https://www.raravind.com/blog/data-science/how-to-pick-the-optimal-image-size-for-training-convolution-neural-network (Accessed: November 10, 2022).

IBISWorld - Industry Market Research, Reports, and Statistics. (n.d.). https://www.ibisworld.com/au/market-size/telehealth/

Interpreting loss curves | machine learning | google developers (2022) Google. Google. Available at: https://developers.google.com/machine-learning/testing-debugging/metrics/interpretic (Accessed: November 10, 2022).

Introduction to convolutional neural networks cnns (2020) AIgents. Available at: https://aigents.co/data-science-blog/publication/introduction-to-convolutional-neural-networks-cnns (Accessed: November 10, 2022).

Janda M, Olsen CM, Mar VJ, Cust AE. Early detection of skin cancer in Australia – current approaches and new opportunities. Public Health Res Pract. 2022;32(1):e3212204.

Kandel, I., & Castelli, M. (2020, May 5). The effect of batch size on the generalizability of the Convolutional Neural Networks on a histopathology dataset. ICT Express. Retrieved November 11, 2022, from https://www.sciencedirect.com/science/article/pii/S2405959519303455

Kaushik, A. (2020, July 21). Understanding ResNet50 architecture. OpenGenus IQ: Computing Expertise & Legacy. https://iq.opengenus.org/resnet50-architecture/

Learn machine learning, artificial intelligence, Business Analytics, data science, Big Data, data visualizations tools and techniques. (no date) Analytics Vidhya. Available at: https://www.analyticsvidhya.com/ (Accessed: November 10, 2022).

Malik, F. (2019, May 20). What are hidden layers? Medium. Retrieved November 11, 2022, from https://medium.com/fintechexplained/what-are-hidden-layers-4f54f7328263#:~:text=One%20hidden%20layer%20is%20sufficient,the%20neural%20network%20can%20solve.

mcc_admin (2022) Melbourne's Skin Cancer Pricing & Skin biopsy cost, Mole Check Clinic. Available at: https://www.molecheck.com.au/pricing-2/ (Accessed: November 12, 2022).

Melanoma - Screening. (2022, June 29). Cancer.Net. https://www.cancer.net/cancer-types/melanoma/screening

Melanoma - Statistics. (2022, June 28). Cancer.Net. https://www.cancer.net/cancer-types/melanoma/statistics

Melanoma mortality rate 5 times higher in developed countries. (2020, November 13). Dermatology Times.

https://www.dermatologytimes.com/view/melanoma-mortality-rate-5-times-higher-developed-countries

Melanoma Treatment (PDQ®)–Patient Version. (2022, September 6). National Cancer Institute. https://www.cancer.gov/types/skin/patient/melanoma-treatment-pdq

Melanoma warning signs and images (2022) The Skin Cancer Foundation. Available at: https://www.skincancer.org/skin-cancer-information/melanoma/melanoma-warning-signs-and-images/ (Accessed: November 10, 2022).

Miiskin: Teledermatology Platform and Skin Tracking App. (n.d.). Miiskin. https://miiskin.com/

MoleScopeTM | Skin screening made simple. (n.d.). Molescope. https://www.molescope.com/

Narein T, A. (2021, October 8). Inception V3 Model Architecture. OpenGenus IQ: Computing Expertise & Legacy. https://iq.opengenus.org/inception-v3-model-architecture/

NCBI - WWW Error Blocked Diagnostic. (n.d.-a). https://www.ncbi.nlm.nih.gov/books/NBK470409/

NCBI - WWW Error Blocked Diagnostic. (n.d.-b). https://pubmed.ncbi.nlm.nih.gov/20197750/

NCBI - WWW Error Blocked Diagnostic. (n.d.-c). https://pubmed.ncbi.nlm.nih.gov/20197750/

Opengenus (2022). Inception V3 Model Architecture. Available at: https://iq.opengenus.org/inception-v3-model-architecture/ (Accessed: November 10, 2022).

Phy, V. (2019) Accuracy is not enough for classification tasks. Available at: https://towardsdatascience.com/accuracy-is-not-enough-for-classification-task-47fca7d6a8ec (Accessed: November 10, 2022).

Ramalingam, A. (2021, June 23). How to pick the optimal image size for training convolution neural network? Medium. Retrieved November 11, 2022, from https://medium.com/analytics-vidhya/how-to-pick-the-optimal-image-size-for-training-convolution-neural-network-65702b880f05

Singh, P. (2021) Confusion matrix: Detailed intuition and trick to learn, Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2021/04/confusion-matrix-detailed-intuition-and-trick-to-learn/ (Accessed: November 10, 2022).

Skin Cancer (Non-Melanoma) - Introduction. (2022, September 20). Cancer.Net. https://www.cancer.net/cancer-types/skin-cancer-non-melanoma/introduction

Skin Cancer Melanoma Detection App. (2022, August 24). SkinVision. https://www.skinvision.com/

Somerville, E. (2021, June 14). Skin cancer check-ups a long time coming as Australia faces huge shortage of dermatologists. ABC News. https://www.abc.net.au/news/2021-06-14/gps-to-help-ease-growing-skin-specialist-waiting-times/100211834

Statista. (2022a, May 12). Smartphone penetration as share of population in Australia 2017-2026. https://www.statista.com/statistics/321477/smartphone-user-penetration-in-australia/

Statista. (2022b, July 27). Internet users as a percentage of the total population Australia 2015-2022. https://www.statista.com/statistics/680142/australia-internet-penetration/

Telehealth Market Size to Hit USD 224.8 Billion by 2030. (n.d.). https://www.precedenceresearch.com/telehealth-market

Telehealth Market Size, Share, Trends & Growth Report, 2030. (n.d.). https://www.grandviewresearch.com/industry-analysis/telehealth-market-report

Tests to diagnose skin cancer (2019) Tests to diagnose | Skin cancer | Cancer Research UK. Available at: https://www.cancerresearchuk.org/about-cancer/skin-cancer/getting-diagnosed/tests-diagnose (Accessed: November 12, 2022).

Thanh-Toan, D., Hoang, T., Victor, P., Yiren, Z., & Zhao, C. (2018, February 26). Accessible Melanoma Detection using Smartphones and Mobile Image Analysis. Arxiv. https://arxiv.org/pdf/1711.09553.pdf

The Skin Cancer Foundation. (2022a, August 17). Melanoma Stages. https://www.skincancer.org/skin-cancer-information/melanoma/the-stages-of-melanoma/

The Skin Cancer Foundation. (2022b, September 14). Melanoma Warning Signs and Images. https://www.skincancer.org/skin-cancer-information/melanoma/melanoma-warning-signs-and-images/

WCRF International. (2022, April 14). Skin cancer statistics | World Cancer Research Fund International. https://www.wcrf.org/cancer-trends/skin-cancer-statistics/

White, L. (2022, February 14). 5 Reasons Why a Good User Interface is Important. CODERSERA. https://codersera.com/blog/why-a-good-user-interface-is-important/