**HOTARU Model Creation Progress Log**Machine Learning Engineer-in-Charge: Nicholas Dale (falconpunch082)

**About**HOTARU (short for Health Observing Technology for Assessing Risks Unveiled) is a convolutional neural network designed to determine from a picture provided by an end-user whether said picture shows signs of skin cancer (defined in this model as melanoma, basal cell carcinoma and squamous cell carcinoma).

HOTARU is accompanied by SPARK (short for Skin Pattern Analysis and Recognition Kit), a pre-processing Python script that serves to analyse a picture by standardising the picture, isolating the skin-spot, and outputting a pre-processed image of the skin-spot to enable more accurate and efficient determination for HOTARU. While HOTARU can run without using SPARK (granted that there is still a barebones preprocessing script that resizes the image into the size HOTARU accepts, 300x300), SPARK is meant to empower HOTARU to predict accurately, considering the risk of misidentification of skin cancer.

**Data Collection and Preprocessing**Image data was collected from the following websites:

* <https://www.kaggle.com/datasets/ahmedxc4/skin-ds>
* <https://www.kaggle.com/datasets/ismailpromus/skin-diseases-image-dataset>
* <https://www.kaggle.com/datasets/subirbiswas19/skin-disease-dataset>
* <https://api.isic-archive.com/collections/70/> (Challenge 2018, 2019 and 2020)

38663 images containing skin cancer and 14669 images containing non-skin cancer features were gathered.

Through a notebook, training, validation, and test datasets were created with the varying splits based on version. In later versions, this notebook is also where the datasets were pre-processed.

**Model Lifecycle**The table below shows the lifecycle of both HOTARU and SPARK, alongside the notebook hosting both programs.

|  |  |  |
| --- | --- | --- |
| HOTARU Version | SPARK Version | Notebook |
| 1.0 | - | production1.ipynb |
| 1.1 | 1.0 | production2.ipynb |
| 1.2 | 1.1 | production3.ipynb |
| 1.21 | 1.1 | production4.ipynb |
| 1.22 | 1.1 | production5.ipynb |
| 1.3 | 1.15 | production6.ipynb |

**HOTARU v1**

Overview

HOTARU v1 was the very first iteration of the CNN model. This model was fed unprocessed images straight from the dataset. Furthermore, SPARK was not introduced at this stage – a preprocessing stage baked into the model was introduced instead.

Model design  
This version featured two sections of the model, which were:

* A Sequential object that encased the CNN, which included:
  + A 16-filter Conv2D input layer with a (3,3) kernel, relu activation and an input\_shape parameter of (300,300,3) followed by a MaxPooling layer,
  + 2 more Conv2D layers (filters were 32 and 16 respectively), each followed by a MaxPooling layer.
  + A Flatten layer for transition from CNN to Dense layers,
  + A Dense layer with 256 neurons and relu activation, and
  + A Dense output layer with one neuron and sigmoid activation (output ranged from 0 to 1, with 0 being non-skin cancer and 1 being skin cancer.), and
* A preprocessing section which included an Input layer of shape (300,300,3) and Resizing layer that converted the input into a size of (300,300).

This model was then compiled and fitted with the following parameters:

* Binary\_crossentropy loss,
* Adam optimiser,
* 20 epochs

Data Preparation  
Training, validation and test datasets were created from the raw dataset in the ratio of 80:10:10 respectively. These photos were reshaped to multi-channel images of size 300x300. No other preprocessing followed.

Training Results  
Upon testing the model with the test dataset, it was determined that the model boasted 80% accuracy, precision of 71.4% and recall of 100%. This model was saved as ‘prod\_1.h5’.

Discussion and Further Action  
Moving forward, it was hypothesised that a preprocessing script would greatly increase the accuracy of the next version as the model would not consider unimportant features like healthy skin and hair when determining whether one had skin cancer or not. Moreover, the preprocessing section of the model was deemed to be redundant as the data entering the model was already sized to be compatible.

**HOTARU v1.1 and SPARK v1.0**

Overview

HOTARU v1.1 was the second iteration of the model. This model also introduced the very first version of SPARK, the preprocessing script. As such, this model was fed with pre-processed images that mostly featured important skin spot features.

Model design  
This version only featured the Sequential object found in HOTARU v1.0; the preprocessing section was removed entirely. No changes were made to the Sequential object, meaning that the CNN in v1.1 is the same as v1.0

This model was then compiled and fitted with the following parameters:

* Binary\_crossentropy loss,
* Adam optimiser,
* 10 epochs

Data Preparation  
Training, validation and test datasets were created from the raw dataset in the ratio of 90:7.5:2.5 respectively. These photos then underwent preprocessing using SPARK.

SPARK v1.0 featured the following stages:

* Resizing the input image into a 300x300 image to preserve information while still being standardised for the model,
* Normalisation of pixel values,
* Creation of a grey-scaled version of the image for thresholding and masking to segment the skin spot from the rest of the skin, and
* Combining the mask and the original image together to create a processed image displaying only the skin-spot (the background was black).

Training results  
Upon testing the model, it boasted an accuracy of 97.2%, with precision of 97.2% and recall of 98.9%. This model was saved as ‘prod\_2.h5’ and ‘prod\_2.pkl’.

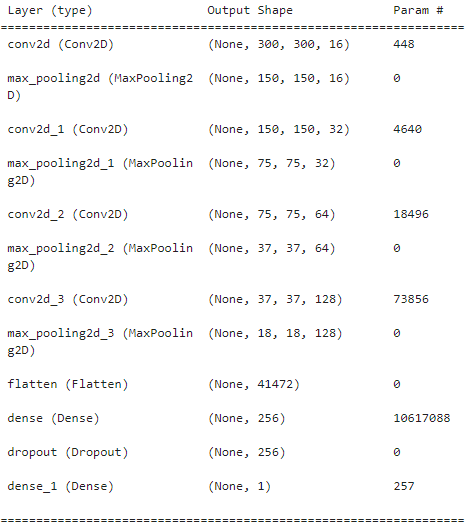
Discussion and Further Action  
Upon reviewing the dataset, it was hypothesised that if the dataset was to be cleaned up so that it only contained pictures of skin (and not body parts like genitals), accuracy would improve as SPARK, and in turn HOTARU, would be more efficient in fulfilling their purpose. In addition, further inspection of the processed photos shows that SPARK v1.0 tended to include body hair into the mask, which resulted in the addition of an unnecessary feature. Moreover, considering that a lot of time was shaved off from training, it is now plausible to add in more layers to HOTARU to make it more accurate without consuming too much time and compute power. Therefore, changes were made to address these hypotheses.

**HOTARU v1.2 and SPARK v1.1**

Overview

HOTARU v1.2 was the third iteration of the model. Changes were made in terms of data preparation and complexity of the model to achieve better real-life accuracy.

Model design  
This version featured improvements to the CNN. It introduced the padding parameter, which was set to ‘same’, to all Conv2D layers. In addition, both the number of Conv2D layers and filters in each Conv2D layer were altered, and another Dense layer alongside a Dropout layer were added at the end of the Sequential object. The output layer remained the same.

The following is a screenshot of the structure of the CNN.

This model was then compiled and fitted with the same parameters as in HOTARU v1.1.

Data Preparation  
Prior to partitioning the raw dataset, the raw dataset was reviewed, and unsatisfactory photos were removed, resulting in the removal of around 11 thousand photos. While the previous versions featured a cancer:non-cancer picture ratio of 2.6:1, this version would learn from a dataset which featured a ratio of 11.7:1 (meaning for every 12 photos with skin cancer featured, there also existed a photo not featuring skin cancer). As such, the disparity between the two classes was exacerbated; the skin cancer class was being oversampled more than before. While this will inevitably result in the appearance of false positives, considering that the use case of this model, it is better to get something checked and find out it is not cancer, than being given a false negative, only to suffer from misidentification later on.

Training, validation and test datasets were created from the improved raw dataset in the ratio of 90:7.5:2.5 respectively. These photos then underwent preprocessing using a new version of SPARK.

SPARK v1.1 featured the same stages as v1.0, except that the masking process was enhanced with a DullRazor hair removal algorithm. In v1.0, the masking process would include body hair, adding unnecessary noise when training the model. This is fixed in v1.1.

Training results  
During the creation of HOTARU v1.2, F1-score was added as an additional metric. In addition, the evaluation now included a confusion matrix and a classification report to further scrutinise v1.2’s performance.

By the end of training, the model provided the following statistics: accuracy of 90-92.1%, with precision of 93.2%, recall of 98.78% and F1-score of 95.9%. When compared to HOTARU v1.1 using the same test dataset (which yielded an accuracy of 90-92.65%, precision of 93.4%, recall of 98.9% and F1-score of 96.1%), it is evident that v1.2 performed slightly worse than v1.1. Furthermore, the confusion matrixes of both versions were nearly identical.

Discussion and Further Action  
The addition of more layers, the cleanup of the raw dataset and the implementation of a new version of SPARK did little to improve the performance of HOTARU, despite all of those actions initially being seen as capable of improving performance. Upon looking at the pre-processed photos, it was observed that some photos were not processed as expected (e.g., the feature was completely left out, leaving a halo of healthy skin). Thus, an attempt was made to improve on SPARK again to hopefully improve HOTARU’s performance.

**HOTARU v1.21**

Overview

HOTARU v1.21 is the fourth iteration of the model made in response to analysing the results gained from v1.2. With this model, a more drastic change was done on the image database so that there were near-equal amounts of skin cancer and non-skin cancer photos in the hopes of allowing for more true results due to the lessened impact of data disparity. It was hypothesised that the immense data disparity caused the model to be so sensitive in detecting skin cancer that any slight resemblance to a feature of skin cancer influenced the decision-making of the model.

Model design  
HOTARU v1.21 followed the same model design as HOTARU v1.2. The only difference between v1.21 and 1.2 was the amount and proportion of data provided.

Due to the small amount of data presented to the model, it was decided that the epochs would be increased from 10 to 100 for the model to have a better chance of learning the differences between the two classes.

Data Preparation  
8.7 thousand photos were provided for this model, compared to 41.9 thousand photos (the number of photos provided for v1.2). 3291 non-skin-cancer pictures and 5453 skin-cancer pictures were at the model’s disposal. All pictures went through preprocessing with SPARK v1.1 after being split into training, validation and test datasets of ratio 80:10:10.

Training results  
Upon the completion of training, the model achieved an accuracy of 55-90%, precision of 91%, recall of 93% and F1-score of 92.6%. In addition, the confusion matrix showed that while the model is now producing more true negatives, it is also producing false positives and negatives.

When comparing HOTARU v1.21 with v1.2, it is very clear that the reduction of image data to reduce the impact of disparate datasets has backfired in that it has caused a significant decrease in performance. The model was not saved.

Discussion and Further Action  
A hypothesis was made that feeding the model with raw data will have an impact on performance. While the difference of performance between HOTARU v1.0 and v1.1 due to the addition of SPARK made it clear that preprocessing images greatly improved accuracy, it was deemed necessary to explore such an avenue to reach a better conclusion on how to move on with the model.

**HOTARU v1.22**

Overview

HOTARU v1.22 was the fifth iteration of the model, which was made to measure the impact of SPARK v1.1 on HOTARU through its absence. In essence, this model was v1.21 but without the influence of the preprocessing script.

Model design  
HOTARU v1.22 followed the same model design as HOTARU v1.21. The only difference was the alteration of epochs from 100 to 20 since during the creation of v1.21, it was observed that by epoch 20 all statistics started to plateau, indicating that any more epochs would be not necessary and would instead waste compute power.

Data Preparation  
8.7 thousand photos were provided for this model, compared to 41.9 thousand photos (the number of photos provided for v1.2). 3291 non-skin-cancer pictures and 5453 skin-cancer pictures were at the model’s disposal. All pictures **DID NOT** go through preprocessing with SPARK v1.1 after being split into training, validation and test datasets of ratio 80:10:10.

Training results  
Upon the completion of training, the model achieved an accuracy of 54-93%, precision of 92.9%, recall of 96.2% and F1-score of 94.5%. In addition, the confusion matrix produced by HOTARU v1.22 and v1.21 are quite similar.

It seemed that with the removal of SPARK from the pipeline, HOTARU was not greatly impacted in terms of performance. In fact, one could say that HOTARU performed slightly better without SPARK when referring to statistics. However, the removal of SPARK did increase the processing time of HOTARU. In v1.21, epochs took a maximum of 98 seconds to complete, and it took 3 seconds for the model to complete its assessment of the test dataset (containing 876 photos). Meanwhile in v1.22, epochs took a maximum of 163 seconds to complete, and it took 16 seconds for HOTARU to finish analysing the test dataset. This proves that SPARK is indeed impactful in improving HOTARU’s performance in terms of processing speed due to its role in removing unnecessary features.

Discussion and Further Action  
Upon looking at the results, two things were certain: that SPARK was influential in increasing HOTARU’s performance but could be improved, and that reducing the dataset to defeat the data disparity issue was not effective in improving HOTARU’s performance. Thus, after some discussion, it was decided to increase the number of non-cancer images through image data augmentation.

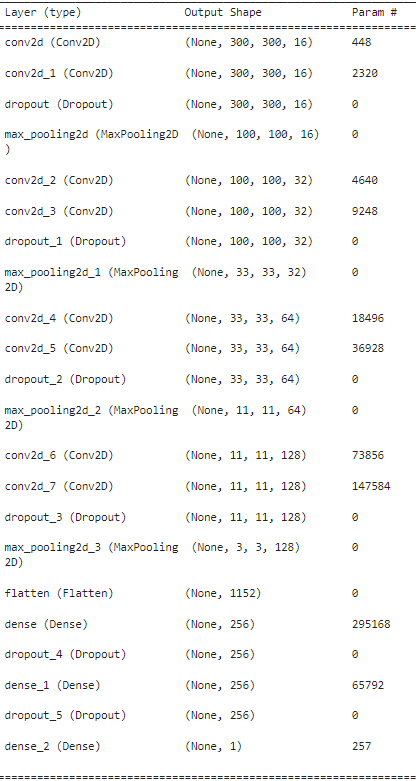
**HOTARU v1.3 and SPARK v1.15**

Overview

HOTARU v1.3 was the sixth iteration of the model. This model aimed to be more accurate than v1.2 by two things:

* being provided with more image data for the non-cancer class through image data augmentation and therefore beating the disparity between both classes of data, which should theoretically improve performance by providing fewer false results, and
* adding more Conv2D layers to improve performance.

Model design  
HOTARU v1.3’s model design is an evolution from v1.2. This model design is a simplified rendition of the VGG16 object detection and classification algorithm. Unlike v1.2, there are now two Conv2D layers with same parameters before a Dropout and MaxPooling2D layer followed. This is repeated a handful of times. Unlike the VGG16 architecture, HOTARU v1.33 only features two Dense layers instead of three, excluding the output layer.

The following is a breakdown of HOTARU v1.3.

The model was then complied with the same parameters as HOTARU v1.2, except that 20 epochs were used instead of 10 to provide more time for the model to learn the differences between both classes.

Data Preparation  
For this version, a new dataset was created by using the dataset used for HOTARU v1.2 and then augmenting every photo in the non-skin-cancer class eleven times to create a larger non-skin-cancer picture database. The augmentations done on each image varied from flipping horizontally, cropping, adding blur, resizing, shifting images and rotation. As a result of the image data augmentation, the number of non-skin-cancer pictures increased from nearly 4 thousand pictures to 39 thousand pictures, which is relatively slightly more than the number of skin-cancer pictures (around 38.6 thousand).

The augmentation code utilised that was not of our making. Please refer to this repository to learn more about the augmentation code: <https://github.com/EdjeElectronics/Image-Augmentation-Examples-for-Machine-Learning>.

All photos were then split into training, validation and test folders in the ratio of 80:10:10 and then processed with SPARK v1.15 before being fed into HOTARU for training. SPARK v1.15 contains the same code as v1.1, except a different contrasting method was used, replacing a previous method, and variables were fine-tuned to reduce the chances of unnecessary features appearing.

Training results  
After training, the model was able to reach an accuracy of 50-92.8%, with precision of 95.1%, recall of 90.3% and F1-score of 92.6%.

When comparing HOTARU v1.3 with v1.2, v1.3 had slightly better accuracy and precision than v1.2, but suffered greatly in recall, resulting in a lower F1-score than v1.2. In addition, it seemed that with the new dataset, the model had a 50:50 chance of making a false result, as seen by the distribution in the confusion matrix and the false positive and negative rates.

Discussion and Further Action  
At this point, despite multiple attempts to improve performance, those attempts proved to be fruitless when using a binary classification framework to assess its performance. Therefore, it is now suggested that instead of using a skin-cancer/not-skin-cancer result utilising a threshold of 0.5, a confidence level is to be provided to the enduser. The model currently outputs a value between 0.00 and 1.00, which can be easily converted to a percentage. This percentage will indicate to the user the chances of the skinspot showing signs of skin cancer, and therefore passes the decision-making process form the model itself to the user. Rather than being given false hope or unnecessary panic upon looking at the results of a unreliable binary classification process, the user is now informed of the chances and is given the opportunity to make a decision to seek medical attention based on the confidence level the model provides.