PROJECT 4 PROPOSAL – GROUP 12

Develop a Machine Learning Model for Melanoma Diagnosis

Contributors: *Ryan James, Lakna Premasinghe, John Porretta, and Praveen Rachakonda*

GitHub (Main) repository link: https://github.com/pkrachakonda/Project4_Gr12.git

Project Goal:

This project aims to develop a predictive model for skin cancer diagnosis for Australian adults aged between 35-50. This will be achieved through the implementation of machine learning (unsupervised, supervised, and Neural Network Algorithms) and train the model from data collected from Australian Health Organisations, scientific experts, and through the Kaggle Database.

Background:

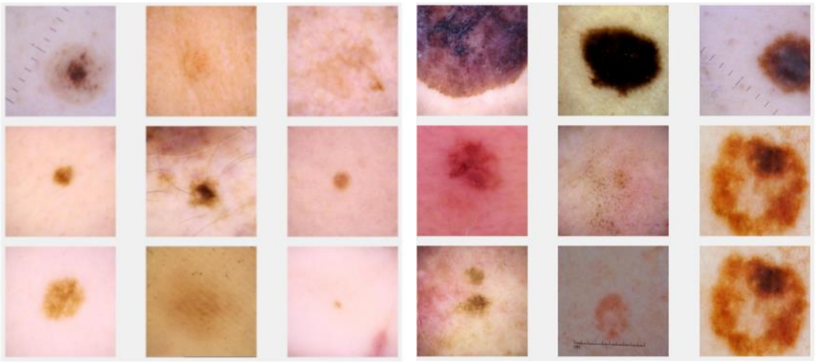
Melanoma is a type of skin cancer that develops in the skin cells. Based on World Health Organisation (WHO), in 2020 Australia ranked second in world for Skin cancer related deaths (4.2 for every 100,000 persons). As per 2023 Australian Institute of Health and Welfare (www.aihw.gov.au) data, Melanoma (Skin cancer) ranks 3 most diagnosed cancer and is found in every 69.4 out of 100,000 Australians (based on 2023 Australian population data). Early detection could play a significant role in the treatment.

Databases:

Proposed datasets to be used as part of this project:

- Cancer Data from the **'Australian Institute of Health and Welfare' (AIHW)**
  - (https://www.aihw.gov.au/reports/cancer/cancer-data-in-australia/data)
  - Specific target group: Australian adults aged between 35-50 years old.
- Harvard Dataverse:
  - The ***HAM1000 Dataset.***
    - Images of common pigmented skin lesions.
  - (https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DBW86T)

- Kaggle
    - Melanoma Tumour Size Predictor
    - (https://www.kaggle.com/datasets/anmolkumar/machine-hack-melanoma-tumor-size-prediction/)



Images of Melanoma from the Kaggle Dataset.

Technologies (Libraries):

- **Web scrapping (Flask APP and Requests):** Download and access the dataset.
- **Scikit-Learn:** Machine Learning functionalities.
- **Python Pandas:** Data manipulation.
- **S3 Bucket/SQL/ SQLAlchemy Database:** Store the data.
- **Python Matplotlib:** Data visualisation.
- **TensorFlow:** Image processing.

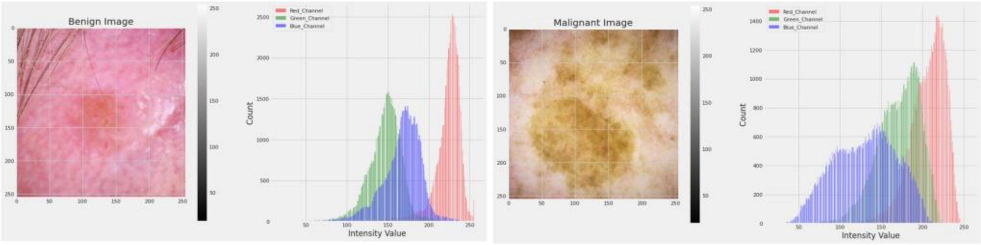
Classification of Cutaneous (skin) Melanoma commonly found for patients 35-50.

- **Superficial Spreading Melanoma:**
    - Most common form (70% approximately).
    - Related to intermittent exposure to the sun.
    - Spotted on the back of female legs and on the backs of males.
    - Characteristics:
        - Colours: tan, brown, grey, black, violaceous, pink and rarely blue or white.
        - Lesion outline: sharply marginated with one or more irregular protrusions/
        - Surface level (skin): Palpable or nodular growth that extends millimetres above the skin level.
- **Nodular Melanoma:**
    - Accounts for a smaller number of melanomas (5%).
    - Located mostly on trunk or limbs of patients.
        - However, occur in the later years of human life.
    - More common in males.
    - Characteristics
        - Colour: brown, black, or blue-black.
        - Lesion outline: Elevated in nature with irregular outlines.
        - Surface level: smooth surface.

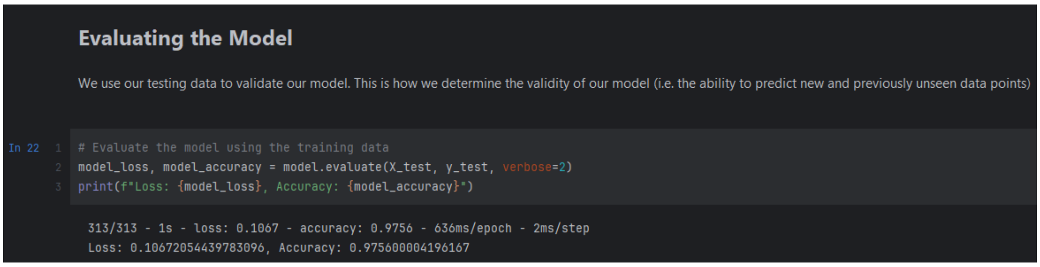
- **Lentigo Maligna Melanoma:**
  - Accounts for 4-15%.
  - Correlates to long-term sun exposure and increasing age.
  - Evolves slowly.
  - Characteristics:
    - Colour: Black, brown, or brown on a tan background.
    - Lesion outline: Tumour is often having irregular outlines and is relatively flat.
    - Surface level: Is located at the neck and head, however, develops in the epidermis level of the skin and takes time to show.

<u>Tentative results (planned)</u>

(a) Histogram showing colour intensity:



(b) Model Predictions



Evaluating the Model

We use our testing data to validate our model. This is how we determine the validity of our model (i.e. the ability to predict new and previously unseen data points)

```
In 22
# Evaluate the model using the training data
model_loss, model_accuracy = model.evaluate(X_test, y_test, verbose=2)
print(f"Loss: {model_loss}, Accuracy: {model_accuracy}")
```

```
313/313 - 1s - loss: 0.1067 - accuracy: 0.9756 - 636ms/epoch - 2ms/step
Loss: 0.10672054439783096, Accuracy: 0.975600004196167
```

<u>Planned Tasks TBA:</u>

Name	Tasks	Due Date
Ryan James		
Lakna Premasinghe		
John Porretta		
Praveen Rachakonda		