HEALTHCARE INDUSTRY
DERMATOLOGY SECTOR

Predicting skin cancer considering severe shortage of the dermatologists to improve the health outcomes

Project Report

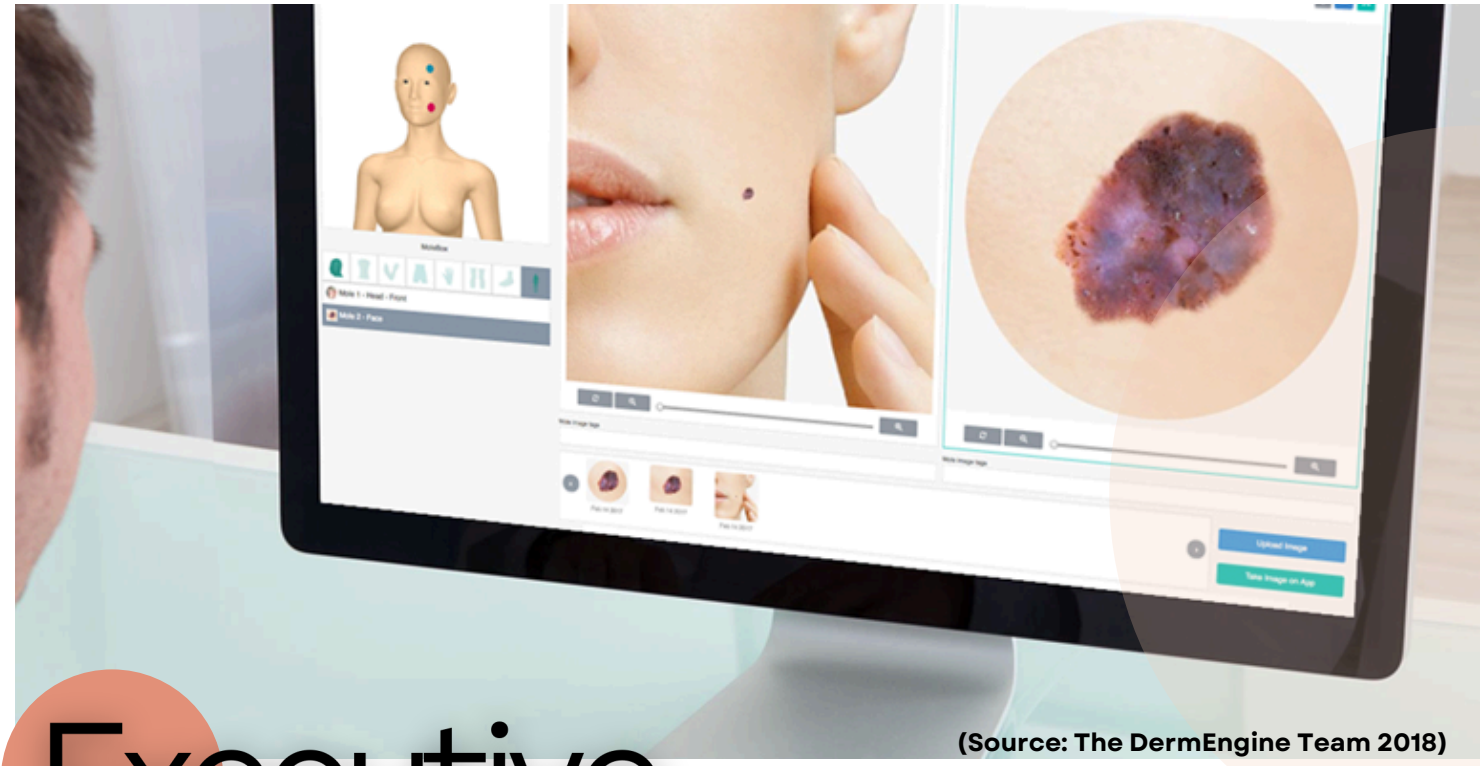
Prepared by: M.H.H. Sevanthi
Student ID 1816338

Kaplan Business School T2 2024
Capstone DATA 6000

For Professor Indu Bala

Table of CONTENTS

Table of CONTENTS



(Source: The DermEngine Team 2018)

Executive Summary

This project aims to address severe shortage of dermatologists in Australia, specifically in regional and rural areas which limits access to diagnosis and treatment for skin cancer patients. "**Only 590 dermatologists are available for 26 million people"**. Author aims to create an image classification system using machine learning to predict skin cancer in patients visiting general practitioners to reduce the long wait times. Analyze 3 types of skin cancer images sourced from 5 platforms (Appendix-6). System identify skin cancer and classify them to skin cancer type. Data set with 277 images used to train two algorithms Neural Network and Random Forest. Neural Network achieved highest precision(69.8%), accuracy(69.9%), and recall (69.9%) rates. Main objective of the project is early diagnosis and to improve diagnosis accuracy of the skin cancer which will reduce the potential life threatening risks and costs. To improve the model's accuracy further could focus on expanding dataset with more diverse skin cancer images and explore advance CNN models specifically designed for image classification.

1.0. INDUSTRY PROBLEM

1.1. Industry background

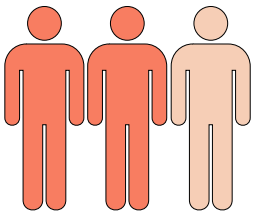
ANNUAL DIAGNOSES IN 2021
550,000+
(mscan 2022)

ECONOMIC BURDEN
$1.2 billion
(mscan 2022)

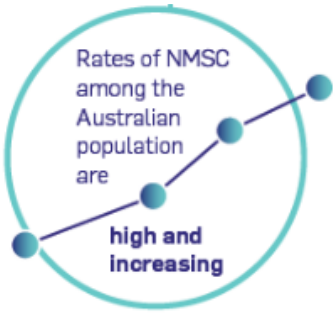
UV EXPOSURE RESPONSIBLE FOR
95%+ skin cancer
(mscan 2022)

80%

Skin cancers account for approximately 80% of all newly diagnosed cancers in Australia each year (Cancer Council 2024).



2 in 3 Australians diagnosed skin cancer by age of 70
(mscan 2022)



(mscan 2022)

Skin Cancer is caused due to DNA damage from UV exposure led to formation of malignant tumors in skin layers **(Appendix 4)**. It's a serious health concern in Australia as evident by the above statistics.

High UV exposure due to its proximity to equator impact for high incident rates which accounts over 95% of skin cancer cases. Depletion of the ozone layer over the Southern Hemisphere further increase the UV exposure. As evident Queensland can be identified as "cancer capital" in Australia due to its demographic characteristics **(Appendix 1).** Australia has taken proactive steps to phase out ozone-depleting.

However, ozone layer is expected to recover even to 1980 level by year 2060 as detailed in (Appendix 2). Beside that other key reasons for high skin cancer incidents are predominant light skin population, outdoor lifestyles, and freckles/ moles in the skin. **Appendix 3 evident** that risk of skin cancer is higher for lighter skin types.

Skin cancer is mainly classified into 3 types: Basal Cell Carcinoma, and Squamous Cell Carcinoma accounts around 99% of cases, and Melanoma (Orazio et al. 2013; **Appendix 4**).

1.2. Business problem

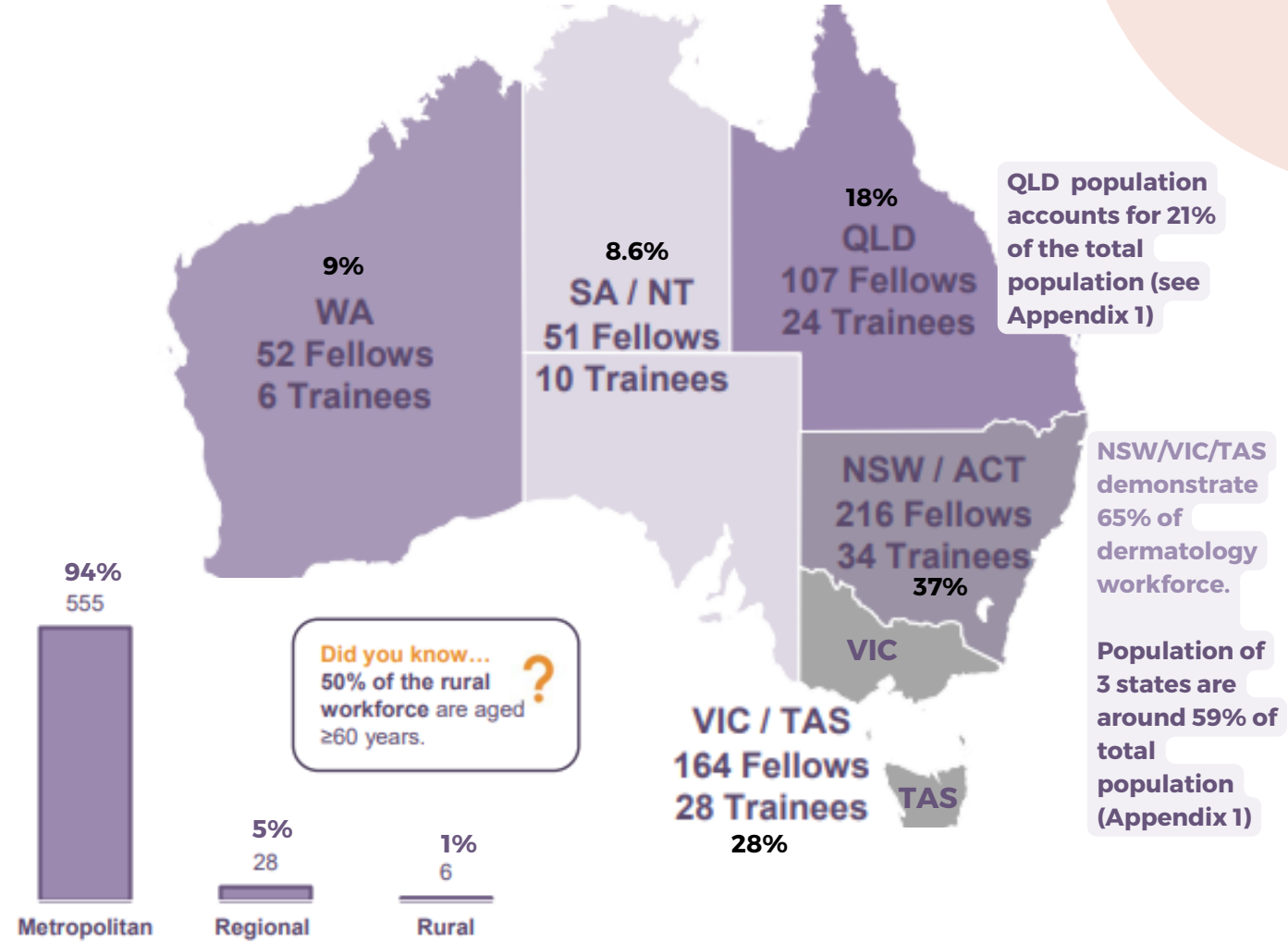


Figure 1: Dermatology workforce Distribution in 2021 (The Australasian College of Dermatologists 2022)

Shortage of dermatalogists

According to Sheppeard (2022) Australia is currently facing severe shortage of dermatologists. According to the above map **only 590 dermatologists are available for 26 million people**. It also highlights a stark contrast between metropolitan, regional, and rural areas. **94% of the workforce are concentrated in metropolitan areas** and **figure 2** evident that **majority are based in capital cites in QLD, NSW, VIC and WA states.**

Figure-1 demonstrate disparity in dermatologists availability across Australia. NSW and VIC states shows highest number of specialists. However, VIC has the lowest skin cancer risk while Queensland state ("Skin cancer capital") demonstrate highest risk, specifically in Brisbane locations **(Appendix 1)**.

Additionally, visual diagnosis accuracy of skin disease is 38% and this accuracy further drops for dark skin tones (Trafton 2024) which highlights a significant gap in dermatology care.

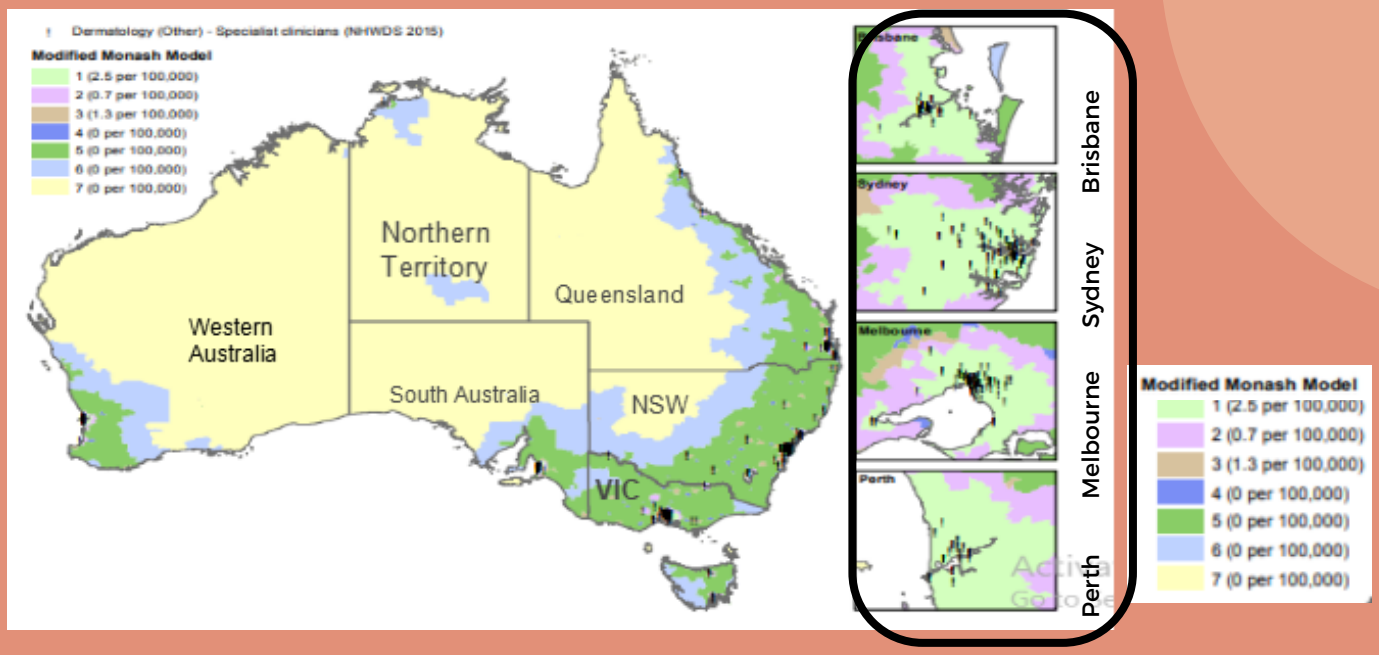


Figure 2: Dermatology workforce Distribution in 2015 in each state (Department of Health 2017)

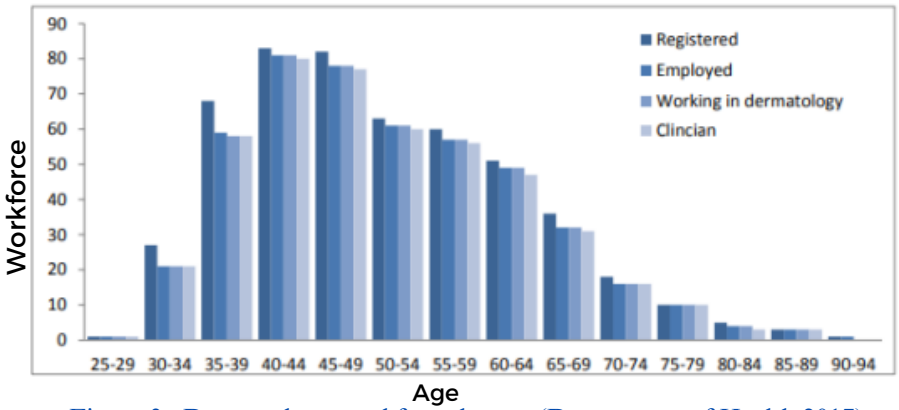


Figure 3 : Dermatology workforce by age (Department of Health 2017)

Furthermore, majority of the workforce **(figure-3)** are ≥50 years.

Figure-4 depicts a decline trend of work hours after 65 years which indicate further reduction of workhours and shrink of the workforce due to potential retirements.

.

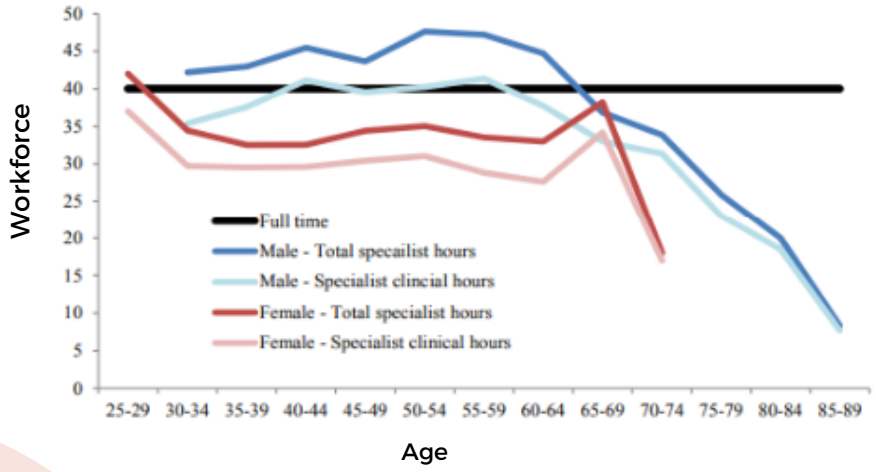


Figure 4 : Skin specialists average hours work per week based on age and gender (Department of Health 2017)

1.3. Importance of solving the business problem

Healthcare providers need efficient tools to accurately and early detection of skin cancer to enhance patients outcomes. Traditional diagnostic methods such as visual diagnosis is time-consuming, with accuracy rates around 38%.

Delayed or misdiagnosed skin cancers specifically aggressive skin cancer like melanoma can be life-threatening and potentially lead to rise of mortality rates and treatment costs. According to Melanoma Research Alliance (2024) prevention and early detection of skin cancer is crucial for successful treatment. However with current shortage of dermatologists makes it challenging.

This project use an innovative approach to skin cancer cancer classification using machine learning model with image data set. Aim to enhance diagnostic accuracy and accessibility, specifically in areas with limited access to dermatology care.

1.4. Formulated business question

Can AI and Machine Learning be used to predict skin cancer to improve patients outcomes?

1.5. Data usage and availability

Integrating diverse data sets from publicly available skin cancer image repositories and medical databases **(Appendix 6)** provided more comprehensive perspective for skin cancer image classification. Diverse datasets covers various types of skin lesions. This provide a robust foundation to train the Random Forest and Neural Network models. Combination of these diverse datasets capture more variations in lesions appearance across dataset populations which enhance the models' capability to generalize and perform accurately on new skin cancer image to help doctors in early detection of skin cancer accurately, assist making patient referrals and treatment plans.

2.0. DATA PROCESSING AND MANAGEMENT

2.1. Data Sources

After extensive research author found five data sets consists of skin cancer image data from Stanford AIMI, the University of Waterloo, Kaggle, and Roboflow to have diverse and quality images of different skin cancer types **(figure 5)**. This is vital to detect skin cancer accurately across different skin cancer types. Author used combined data set of 1038 images classified into 3 classes : Melanoma, Basal Cell Carcinoma (BCC), and Squamous Cell Carcinoma (SCC). All the images are label according to the skin cancer type. **Kindly refer Appendix 6 for original data set links and clean data set link**.

2.1.1. Image data

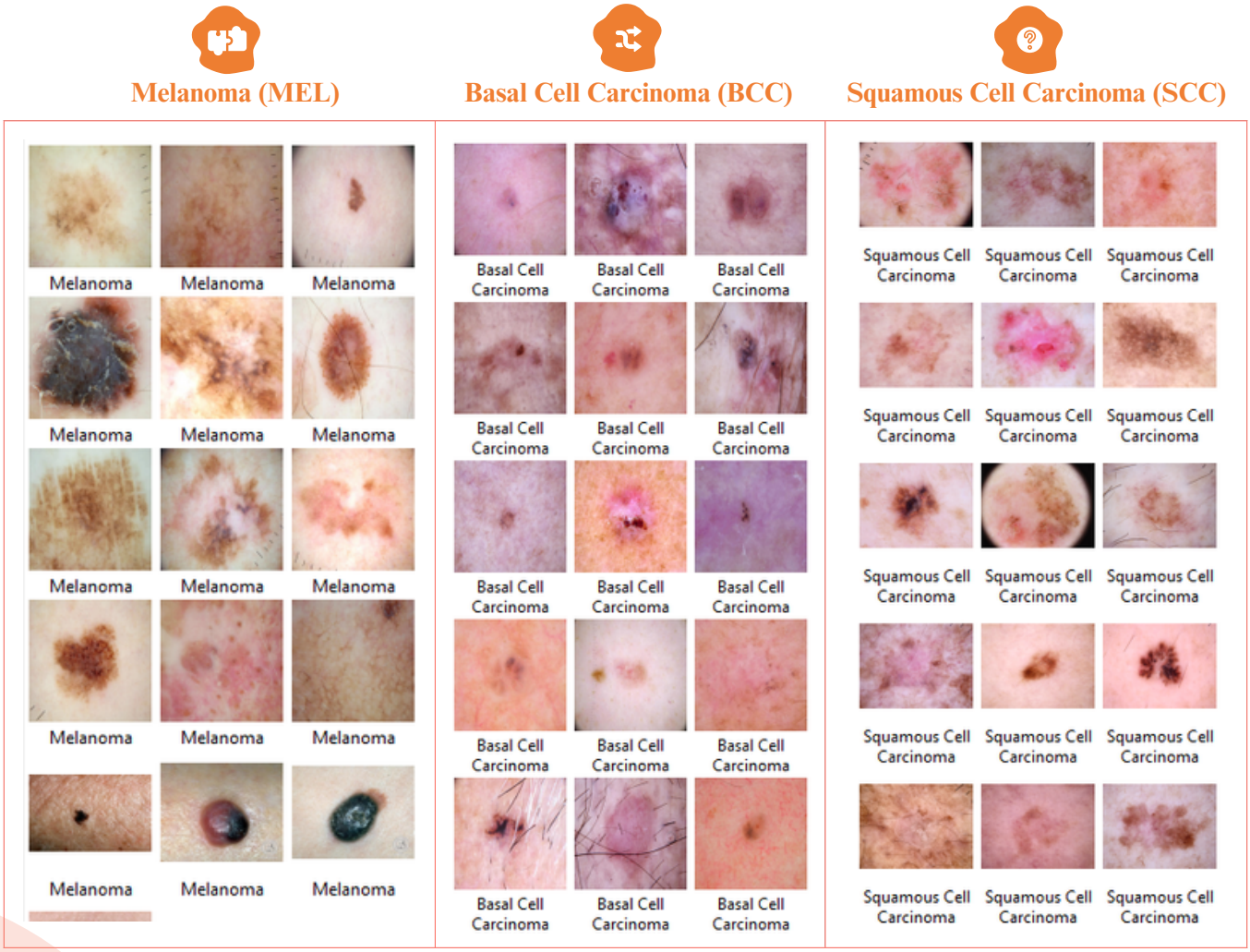


Figure 5: Image data-set for skin cancer (Source: Stanford AIMI, the university of Waterloo, Kaggle, and Roboflow)

2.2. Descriptive analytics

The author used hierarchical clustering unsupervised machine learning algorithm as a exploratory data analytics tool to obtain deep insights into the data before using supervised models. Dendrograms visualize how data is naturally group (Diaz-Maurin & Giampietro 2013) together (figure 7) without pre-defined number of clusters (Coursera 2024). It helps to detect outliers before training the model to identify images that do not fit within any cluster in the data set. Outliers may negatively impact on predictive models. Therefore, addressing them before it fed into the model improve the data quality and performance of the models to support automating skin cancer detection using AI.

2.3. Predictive analytics

Predictive analysis used to predict the data outcomes by learning from the current and historical data (Cote 2021). This project used Random Forest and Neural Networks models to predict skin cancer accurately. General practitioners can diagnose skin cancer at early stage and able to save time and costs since this system able to categorize skin cancer type accurately.

2.4. Overview of the data cleaning, preparation, and mining

Author was collected more than 2300 skin cancer image data. After analyzing the data set author excluded other skin diseases folders and deleted the duplicated images. Only 1038 images were retained and used for this project and ensured images were properly labeled.

Used descriptive analytics (table) to understand the classes were imbalanced **(Appendix 7)**. It was identified that lower number of data among each data class is 346**.** Then used that number to split data to training (80%) and testing (20%). 69 images were use as testing data and 277 images used as training data in each class **(Appendix 7- data combination and pre-processing)**. Project focus to have balance data set to prevent biasness and to ensure a good representation of the images from each data set.

Finally, import the cleaned and balanced image to the orange software to further analysis such as exploratory analysis to ensure most impactful features for predictive analysis as explained in 2.2 descriptive analysis section.

3.0. DATA ANALYTICS METHODOLOGY

In this project orange data mining software is used due to its versatility to use multiple machine learning models in parallel to assess the best performance model.

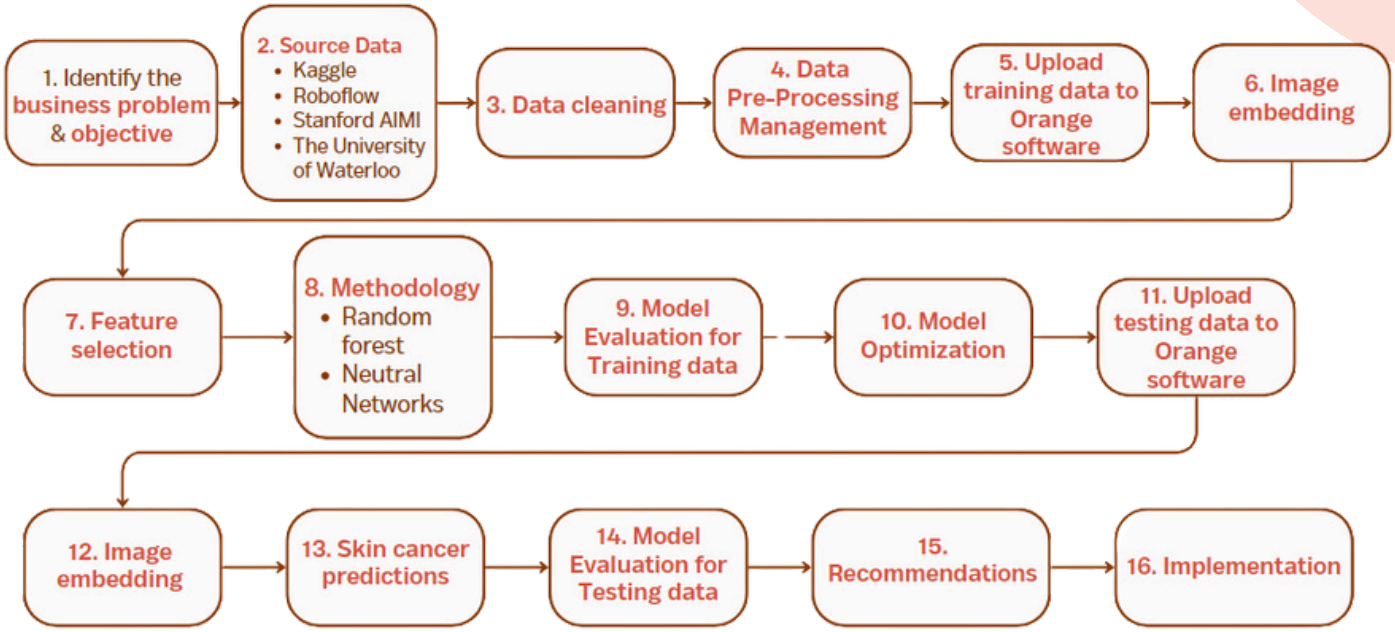


Figure 6 : Flow chart for the data analytics methodology (Author developed)

This flowchart was created by the author for the methodology in skin cancer prediction project. After identifying the importance of developing a tool for healthcare sector to help early detection of skin cancer to improve patients outcomes, author begin the process with data collection, pre-processing and management (Appendix 6 & 7) which is vital for the project success. Then training dataset images was uploaded to the orange data mining software using "**import widget"(5)**. (Step 6) For machine learning model to understand the images need to convert it to vector representation to read using **"Image embedded"** widget (Appendix 9). **"Select column"** widget used to select the image label as major feature to train machine-learning models (7). Author selected 4 models to train the model (Appendix 9), however selected Neural Networks and Random Forest for the project.

Neural Network model selected due to its ability to learn from complex data through layered neurons and handle complex data sets (Masood & Ahmad 2021). This ability is vital for this skin cancer project due to the diverse features present in our skin cancer data set. Also data set is consists with more than 1000 images. This model excel learning hierarchical features where each layer lean continuously more abstract representation of data which makes specifically suitable for image classification (Altexsoft 2019).

Random Forest also used because it has the ability to handle classified data sets to provide more accurate predictions (IBM 2023). Its ensemble approach help to construct multiple decision trees for large data set use in this project. Model is robust against overfitting which is important for complex skin cancer lesion data set. Model work well with NN due to its ability in handling high-dimensional data without needing extensive feature selection. Models will provide more robust and insightful results.

Then performance metrics were evaluated for the models (9), and fine tune the models to obtain better results (10). For instance, structured Neural Networks and Random Forest models (Appendix 9) through hyperparameter tuning aiming to learn hierarchical features more effectively by using its flexibility to handle large datasets and its superior performance in similar medical imaging tasks. After achieving the best performance results with training dataset same process was used to testing data (11, 12, 13, 14).

Finally, communicate the results and recommendations for further analysis (15) and implement the project in the real world medical settings (16)

4.0. VISUALISATION AND EVALUATION OF RESULTS

4.1. Descriptive analytics

The four dendrograms (figure-7) visually shows how skin images are grouped based on the feature similarities. These clusters visualize valuable insights to identify natural relationships between different skin cancer types before applying to predictive models. It is evident that the images belongs to the same cancer type often group together demonstrating natural relationship. Also, it demonstrate images from different cancer types group together based on the feature similarities across cancer types. It helps to detect patterns, identify potential outliers and assess class separability by providing insights how accurately predictive models could distinguish skin cancer types.

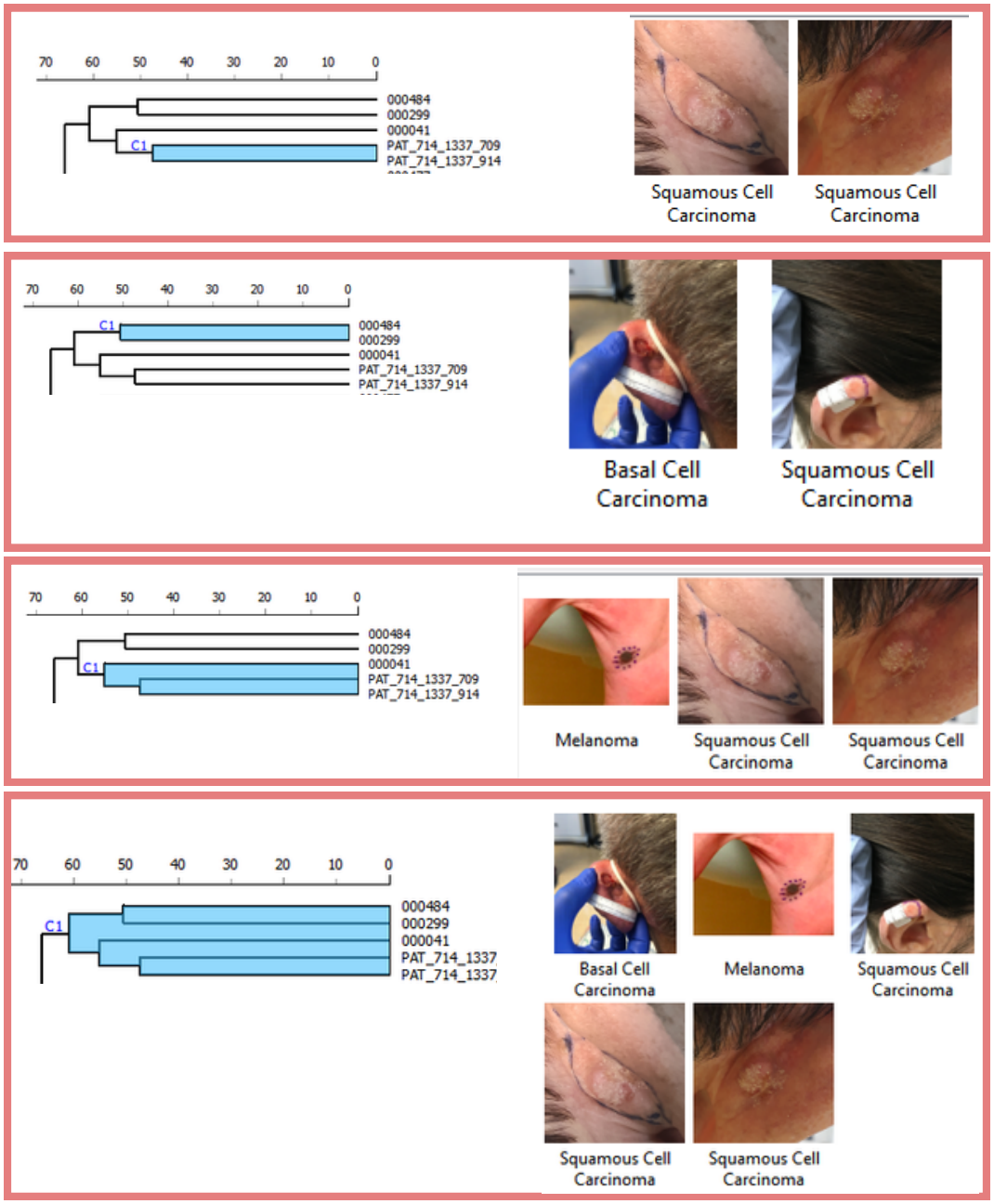


Figure 7: Visualization of hierarchical clustering for skin cancer (Orange data mining software)

4.2. Predictive analytics

Training data

Results shows the performance of four models on skin cancer data set and evident that Neural Network (NN) model highest Accuracy (0.849), precision (0.698), and recall (0.698). High accuracy indicate that it has strong capability to differentiate between skin cancer classes. Below ROC curves (figure 9) reinforce that demonstrating better prediction (shows most distance from diagonal line).

However, recall (69.9%) is significantly important due to high cost of incorrect diagnosis and risk of miss identification of skin cancer type. This indicates that NN model failed to identify 30% of skin cancer types correctly.

Model	AUC	CA	F1	Prec	Recall	MCC
Random Forest	0.798	0.610	0.610	0.610	0.610	0.415
kNN	0.784	0.611	0.602	0.620	0.611	0.427
Logistic Regression	0.834	0.681	0.681	0.681	0.681	0.522
Neural Network	0.849	0.699	0.699	0.698	0.699	0.549

Figure 8 : Performance metrics for training data

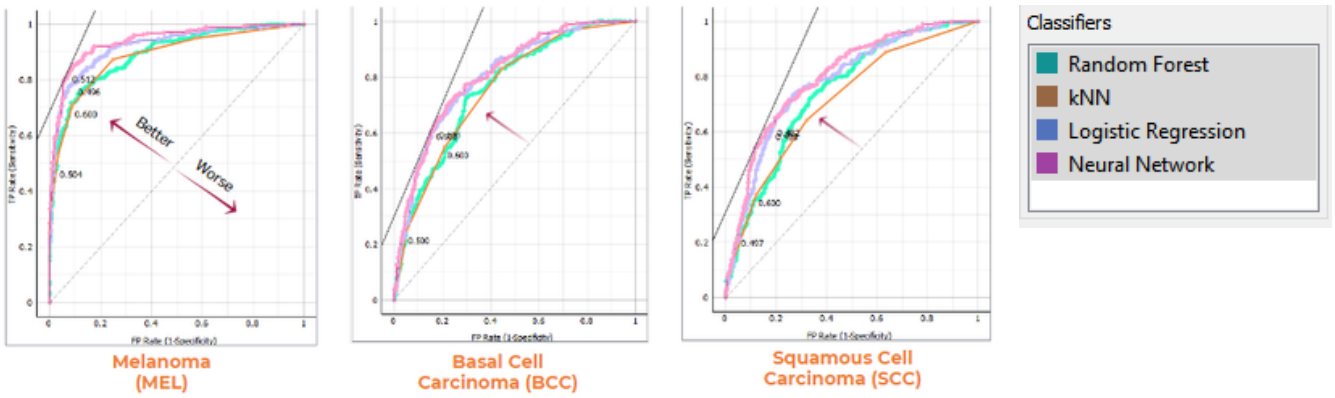


Figure 9 : ROC curve analysis for training data

Training data

Figure 10 and 11 shows Random Forest and Neural Network models demonstrate many number of correctly identified skin cancer types (true positive and true negative). However, considerable number of images were misclassified. For a skin cancer classification diagnose project these misclassifications is a significant issue. Testing data demonstrate (please refer **Appendix 10)** shows slightly fewer performance metrics than training data.

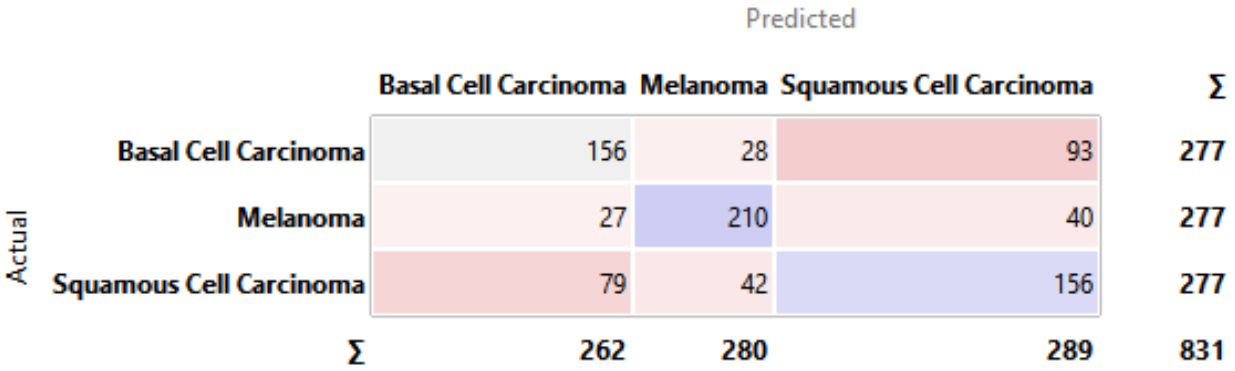


Figure 10: Confusion Matrix - Random forest (Skin cancer dataset)

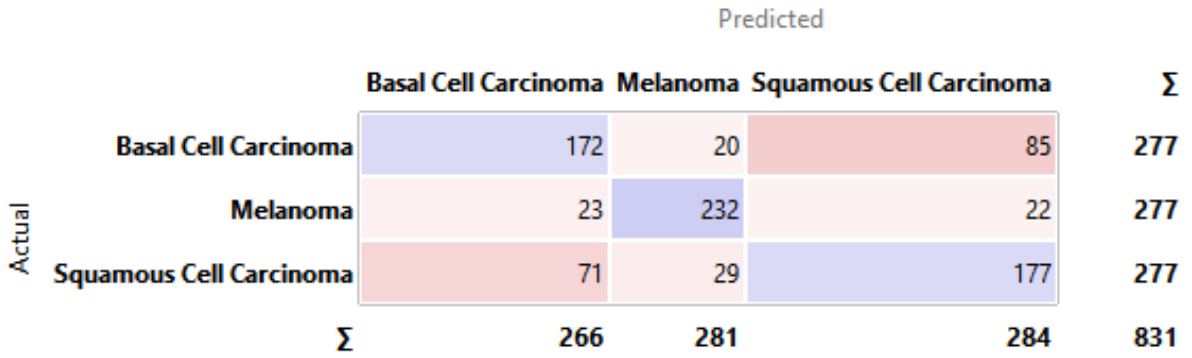


Figure 11: Confusion Matrix - Neural Network (Skin cancer dataset)

4.3. Reflection of the efficacy of the techniques/software

Although Neural Network shows high accuracy (84.9%), a considerable number of images were misclassified. Recall rate indicate that model failed to classify 30% skin cancer cases among three types correctly demonstrate the challenge to implement in the real world medical settings. Misclassification between cancer types will lead to serious issues. So, required improvement even the model could be a helpful tool for doctors.

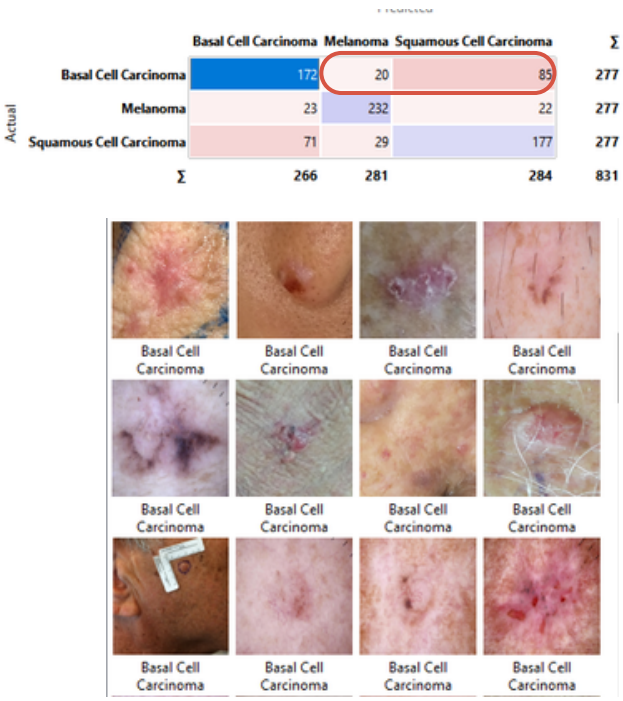


Figure 12: Incorrectly classified BCC cases

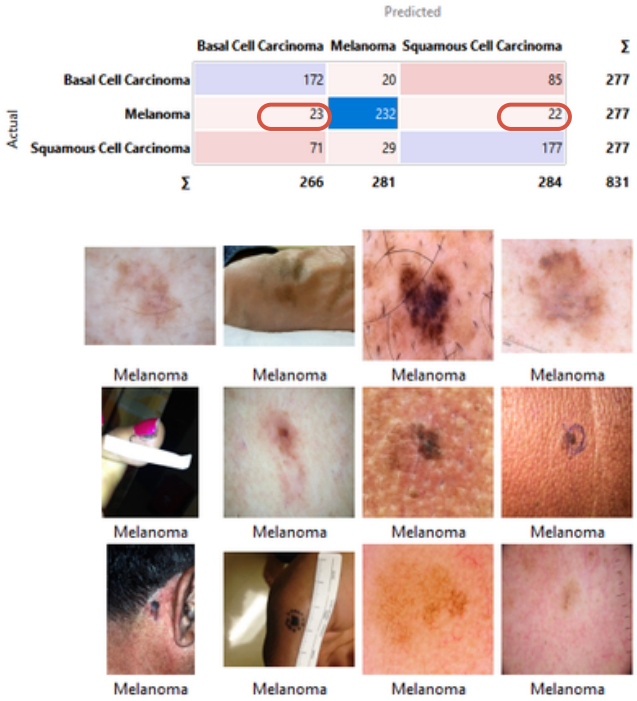


Figure 13: Incorrectly classified melanoma cases

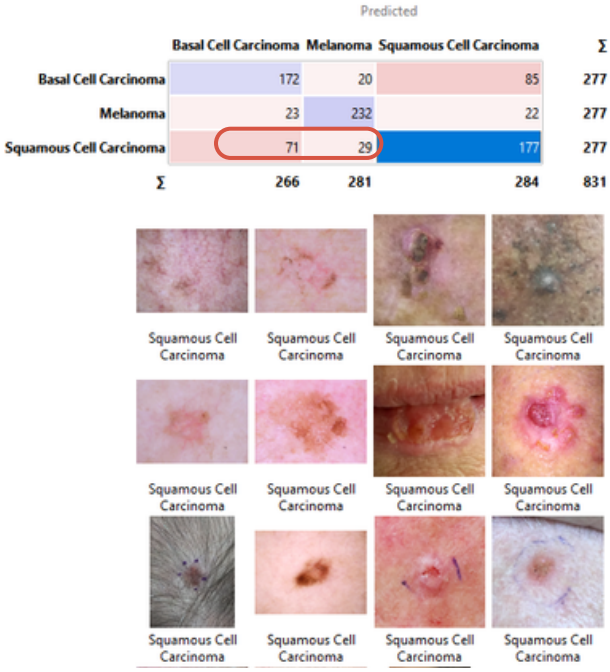


Figure 14: Incorrectly classified SCC cases

Orange software is used for the project due to no requirements for coding, user friendly interface, and the software ability to compare different algorithms in one place. With the workflow if adjustment made data will update automatically in all steps which makes easy to use.

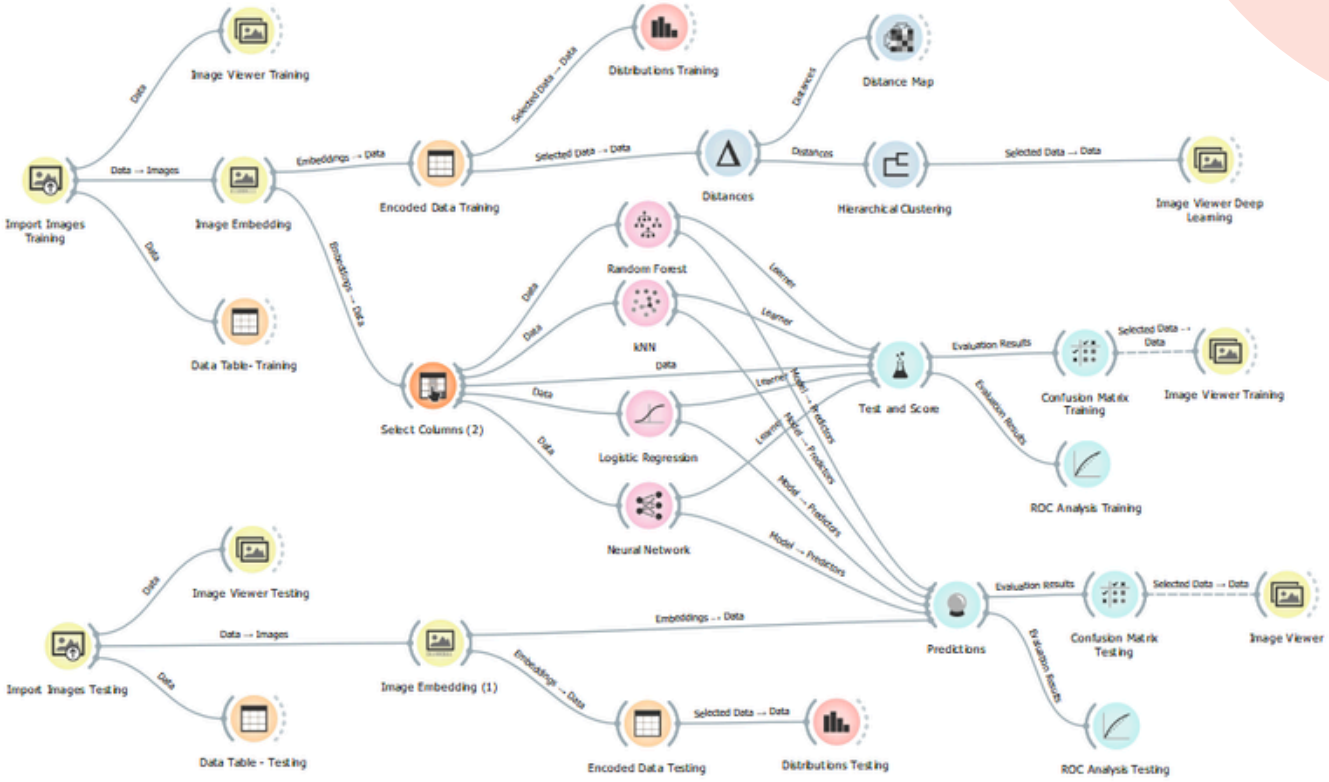


Figure 15: Orange Workflow (Orange data mining software)

5.0. RECOMMENDATIONS

5.1. Results and Insights

Purpose of this project is to provide innovative approach to classify skin cancer using orange data mining tool for skin cancer image classification. Neural Network model achieved highest accuracy (84.9%) based on training data. This tool can implement in clinics and hospitals to assist doctors classifying three types of skin cancer and it would help doctors to make informed decisions and save time and cost. However, for skin cancer project accurate classification is vital. Precision and recall rates also plays a vital role. NN model misclassify around 30% of skin cancer types (recall rate 69.9%) which highlight the challenge in implementation in real world. Misclassifications will results suboptimal and delayed treatments or may worsen patients situation which will impact on patients health outcomes. Melanoma skin cancer require urgent treatments and early detection, it is vital to improve survival rates (Appendix 4). So it is vital to improve the model's recall and precision rates since misclassification will further worse health condition.

5.2. Limitations

The following limitations are required to address

- Two dataset author used included diverse skin tones. However, could enhance the diversity of the dataset by having balance representation of all skin cancer types, gender, and lesion characteristics to enhance model's ability to generalize data.

- Explore more advance analytic models such as Convolutional Neural Networks (CNN) which has ability to capture even local spatial relationships in image, which is more specific for image classification. It potentially improve model's capability in capturing stubble differences between skin cancer types. NN is lack of ability to capture spatial relationships (Mikhailiuk 2022).

- Use adaptive learning method which allow the model to improve iteratively based on dermatologists feedbacks and new data which ensures continued accuracy and relevance in distinguishing between skin cancer types.

5.3. Future suggestions

In addition to skin cancer classification, project should also can enhance scope considering to develop a system to detect and track skin cancer lesions over time. This will be vital for monitoring the progression of each cancer type.

This project also could explore integrating genetic data and history details along with image analysis to provide more comprehensive approach to assess cancer risk and to provide personalized treatment planning for each cancer type.

6.0. DATA ETHICS AND SECURITY

6.1. Privacy, legal, security, and ethical considerations

Author conduct this project for academic purpose using publicly available data by adhering to data ethical standards and protection principles. Therefore, legal, privacy and security concerns were inherently covered by using open source data.

6.2. Reflection on the accuracy and transparency of visualizations

Combined dataset is created in the data pre-processing phase and described in the data processing chapter, focusing on maintaining a balance skin cancer data classes and image origins to avoid biasness. After that author developed model tuning by ensuring transparency and academic integrity.

6.3. Recommendations for the future analytics

- In real-life situations, a new dataset is required to create considering ethical factors including data privacy. Information collected from patients should be anonymous, unless they gave consent to use and distribute data.
- Create guidelines to use AI responsibly in medical diagnosis.
- Conduct ongoing discussions related to decision making process in skin cancer prediction system
- Integrate human evaluations with machine learning outcomes
- Periodical maintenance and testing.

7.0. REFERENCES

- Altexsoft 2019, *Image Recognition with Deep Neural Networks and its Use Cases*, AltexSoft, viewed 30 September 2024, <https://www.altexsoft.com/blog/image-recognition-neural-networks-use-cases/>.

- Anthony.Augustine 2023, *Melanoma in Men vs Women: Differences in Risk & Treatment,* SunDoctors, viewed 1 October 2024, <https://sundoctors.com.au/blog/differences-in-melanoma-men-vs-women/>.

- Australian Bureau of Statistics 2024, *National, state and territory population*, www.abs.gov.au, viewed 16 September 2024, <https://www.abs.gov.au/statistics/people/population/national-state-and-territory-population/latest-release>.

- Australian Government Department of Climate Change, Energy, the Environment and Water 2023, *Outcomes of the Review of the Ozone Protection and Synthetic Greenhouse Gas Management Program*, viewed 15 September 2024, <https://www.dcceew.gov.au/environment/protection/ozone/publications/factsheet-opsggm-review-outcomes>.

- Australian Government Department of Climate Change, Energy, the Environment and Water 2024, *Montreal Protocol on Substances that Deplete the Ozone Layer*, Australian Government Department of Climate Change, Energy, the Environment and Water, viewed 12 September 2024, <https://www.dcceew.gov.au/environment/protection/ozone/montreal-protocol#:~:text=As%20one%20of%20the%20first,out%20of%20ozone%20depleting%20substances.>.

- Australian Institute of Health and Welfare 2024, *Non-melanoma skin cancer: general practice consultations, hospitalisation and mortality, Summary*, Australian Institute of Health and Welfare, viewed 1 October 2024, <https://www.aihw.gov.au/reports/cancer/non-melanoma-skin-cancer/summary>.

- Better Health Channel 2014, *Melanoma*, Vic.gov.au, viewed 14 September 2024, <https://www.betterhealth.vic.gov.au/health/conditionsandtreatments/melanoma>.

- Caliper 2024, *Australia Health Care Mapping*, www.caliper.com, viewed 16 September 2024, <https://www.caliper.com/maptitude/publichealth/health-care-mapping-software-australia.htm>.

- Cancer Council 2023, *Incidence and mortality | National Cancer Prevention Policy Skin Cancer Statistics and Issues*, www.cancer.org.au, viewed 20 September 2024, <https://www.cancer.org.au/about-us/policy-and-advocacy/prevention/uv-radiation/related-resources/skin-cancer-incidence-and-mortality>.

- Cancer Council 2024, *Basal and Squamous Cell Carcinoma | Non-melanoma skin cancer*, Cancer.org.au, viewed 16 September 2024, <https://cancer.org.au/cancer-information/types-of-cancer/non-melanoma-skin-cancer#:~:text=Non%2Dmelanoma%20skin%20cancers%2C%20now>.

- Cancer Council NSW 2021, *About skin cancer*, Cancer Council NSW, viewed 14 September 2024, <https://www.cancercouncil.com.au/skin-cancer/about-skin-cancer/>.

- Cancer Research UK 2020, *Symptoms of Advanced Melanoma | Melanoma | Cancer Research UK,* Cancerresearchuk.org, Cancer Research UK, viewed 1 October 2024, <https://www.cancerresearchuk.org/about-cancer/melanoma/advanced-melanoma/symptoms-advanced-melanoma>.

- Chen, JJ 2003, 'COMMUNICATING COMPLEX INFORMATION: THE INTERPRETATION OF STATISTICAL INTERACTION IN MULTIPLE LOGISTIC REGRESSION ANALYSIS', *American Journal of Public Health*, vol. 93, no. 9, pp. 1376-a-1377, viewed 27 September 2024, <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1447969/>.

- Ciążyńska, M, Kamińska-Winciorek, G, Lange, D, Lewandowski, B, Reich, A, Sławińska, M, Pabianek, M, Szczepaniak, K, Hankiewicz, A, Ułańska, M, Morawiec, J, Błasińska-Morawiec, M, Morawiec, Z, Piekarski, J, Nejc, D, Brodowski, R, Zaryczańska, A, Sobjanek, M, Nowicki, RJ & Owczarek, W 2021, 'The incidence and clinical analysis of non-melanoma skin cancer', *Scientific Reports*, vol. 11, no. 1, viewed 1 October 2024, <https://www.nature.com/articles/s41598-021-83502-8>.

- Cote, C 2021, *What Is Predictive Analytics? 5 Examples* | HBS Online, Business Insights - Blog, Harvard Business School, viewed 2 October 2024, <https://online.hbs.edu/blog/post/predictive-analytics>.

- Coursera 2024, *What Is Hierarchical Clustering?*, Coursera, viewed 1 October 2024, <https://www.coursera.org/articles/hierarchical-clustering>.

- Department of the Environment and Heritage 2001, *Management Strategy Chlorofluorocarbon, Australian Government Department of the Environment and Heritage*, viewed 15 September 2024, <https://www.dcceew.gov.au/sites/default/files/documents/cfcms.pdf>.

- Diaz-Maurin, F & Giampietro, M 2013, *Dendrogram - an overview* | ScienceDirect Topics, www.sciencedirect.com, viewed 2 October 2024, <*https://www.sciencedirect.com/topics/agricultural-and-biological-sciences/dendrogram*>.

- Elliott, TM, Whiteman, DC, Olsen, CM & Gordon, LG 2017, 'Estimated Healthcare Costs of Melanoma in Australia Over 3 Years Post-Diagnosis', *Applied health economics and health policy*, vol. 15, New Zealand, no. 6, pp. 805–816, viewed 16 September 2024, <https://www.ncbi.nlm.nih.gov/pubmed/28756584>.

- Gruber, P, Shah, M & Zito, PM 2020, *Skin Cancer*, PubMed, StatPearls Publishing, Treasure Island (FL), viewed 13 September 2024, <https://www.ncbi.nlm.nih.gov/books/NBK441949/>.

- Healthdirect Australia 2019, *Skin cancer and melanomas*, www.healthdirect.gov.au, viewed 15 September 2024, <https://www.healthdirect.gov.au/skin-cancer-and-melanomas>.

- IBM 2023, *What Is Random Forest?* | IBM, www.ibm.com, IBM, viewed 2 October 2024, <https://www.ibm.com/topics/random-forest>.

- Johnson, S 2018, *Cancer capitals of Australia are revealed with some surprising results*, Mail Online, viewed 16 September 2024, <https://www.dailymail.co.uk/news/article-6203943/The-cancer-capitals-Australia-revealed-new-Cancer-Council-Queensland-interactive-map.html>.

- Kuhn, M & Johnson, K 2013, *Applied Predictive Modeling,* viewed 25 September 2024, <https://www.ic.unicamp.br/~wainer/cursos/1s2021/432/2013_Book_AppliedPredictiveModeling.pdf>.

- Masood, A & Ahmad, K 2021, *Deep Neural Network - an overview* | ScienceDirect Topics, www.sciencedirect.com, viewed 3 October 2024, <https://www.sciencedirect.com/topics/engineering/deep-neural-network>.

- Melanoma Institute Australia 2023, *Melanoma Facts*, Melanoma Institute Australia, viewed 16 September 2024, <https://melanoma.org.au/about-melanoma/melanoma-facts/>.

- Melanoma Research Alliance 2024, *Melanoma Early Detection: What You Need to Know,* Melanoma Research Alliance, viewed 1 October 2024, <*https://www.curemelanoma.org/about-melanoma/educate-yourself*>.

- Mikhailiuk, A 2022, *Convolutional neural networks — the essential summary*, Medium, viewed 1 October 2024, <https://towardsdatascience.com/cnn-cheat-sheet-the-essential-summary-for-a-quick-start-58820a14d3b4>.

- mscan 2022, Facts and figures, MSCAN, viewed 21 August 2024, <https://mscan.org.au/learning-hub/skin-cancer/facts-and-figures/>.

- NOAA Chemical Sciences Laboratory (CSL) 2020, NOAA CSL: 2020 *News & Events: World Ozone Day 2020 marks 35 years of ozone layer protection*, Noaa.gov, viewed 20 September 2024, <https://csl.noaa.gov/news/2020/289_0916.html>.

- Orazio, JD, Jarrett, S, Amaro-Ortiz, A & Scott, T 2013, 'UV radiation and the skin', *International Journal of Molecular Sciences,* vol. 14, no. 6, pp. 12222–12248, viewed 19 September 2024, <https://pubmed.ncbi.nlm.nih.gov/23749111/>.

- The Australasian College of Dermatologists 2022, *Facebook,* Facebook.com, viewed 21 September 2024, <https://www.facebook.com/photo/?fbid=2231831810365905&set=a.1671739233041835>.

- The DermEngine Team 2018, *Top Ways AI Is Revolutionizing Dermatology,* Dermengine.com, viewed 6 September 2024, <https://www.dermengine.com/blog/artificial-intelligence-ai-dermoscopy>.

- Trafton, A 2024, *Doctors have more difficulty diagnosing disease when looking at images of darker skin*, MIT News | Massachusetts Institute of Technology, viewed 10 September 2024, <https://news.mit.edu/2024/doctors-more-difficulty-diagnosing-diseases-images-darker-skin-0205>.

- UN environment programme 2023, *Ozone layer recovery is on track, helping avoid global warming by 0.5°C*, UN Environment, viewed 15 September 2024, <https://www.unep.org/news-and-stories/press-release/ozone-layer-recovery-track-helping-avoid-global-warming-05degc#:~:text=On%20track%20to%20full%20recovery&text=The%20Montreal%20Protocol%20has%20thus>.

- Western, LM, Vollmer, MK, Krummel, PB, Adcock, KE, Fraser, PJ, Harth, CM, Langenfelds, RL, Montzka, SA, Mühle, J, O'Doherty, S, Oram, DE, Reimann, S, Rigby, M, Vimont, I, Weiss, RF, Young, D & Laube, JC 2023, 'Global increase of ozone-depleting chlorofluorocarbons from 2010 to 2020', *Nature Geoscience*, vol. 16, pp. 1–5, viewed 14 September 2024, <*https://www.nature.com/articles/s41561-023-01147-w*>.

8. APPENDIX

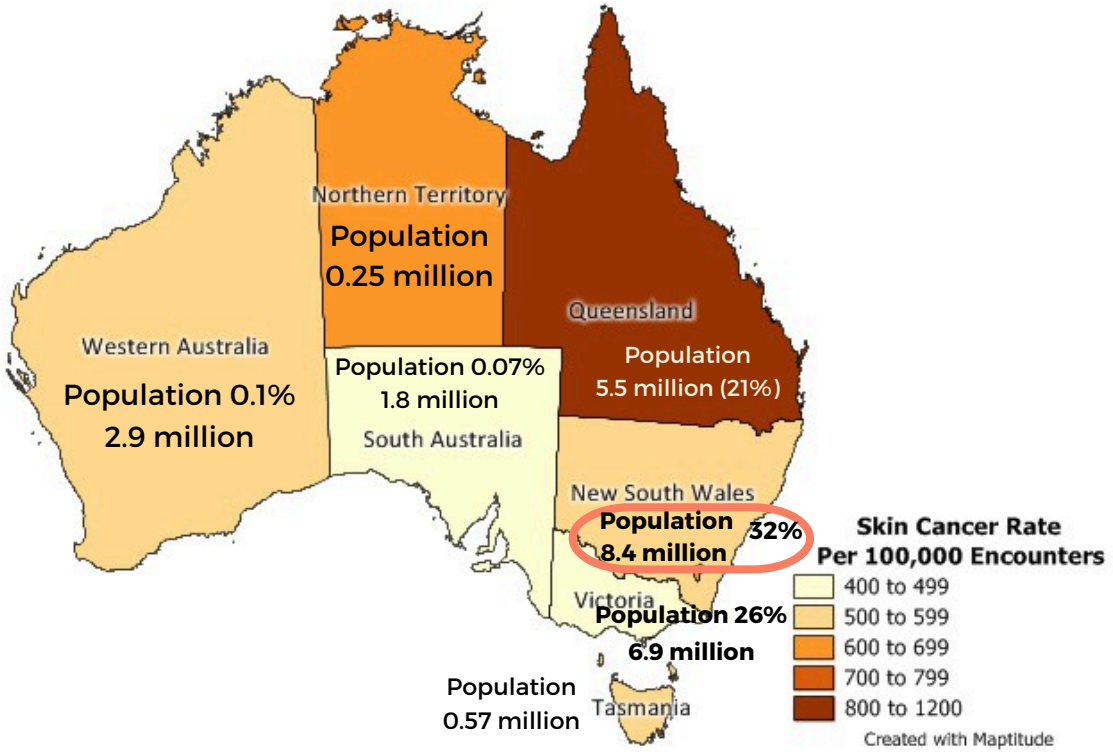


Figure 16: Skin cancer risk in each state (Caliper 2024; Australian Bureau of Statistics 2024)

STATE WISE SKIN CANCER REVIEW

Figure 1 shows skin cancer rates across Australian regions along with the population by demonstrating significant variations of the regions in different states. It highlights in dark shade **Queensland with 5.5 million population as the highest skin cancer rates state**. **Northern Territory** has the second highest skin cancer rates with lowest population. In contrast Victoria region demonstrate lower incidence rates despite second large population.

Figure 2 demonstrate the skin cancer hotspots such as Byron Bay, which has 146% skin cancer rate (Johnson 2018). Also north coast of New South Wales and south-east Queensland which are known for popular beaches are considered as high risk locations. This evidence the demography has high impact on skin cancer.

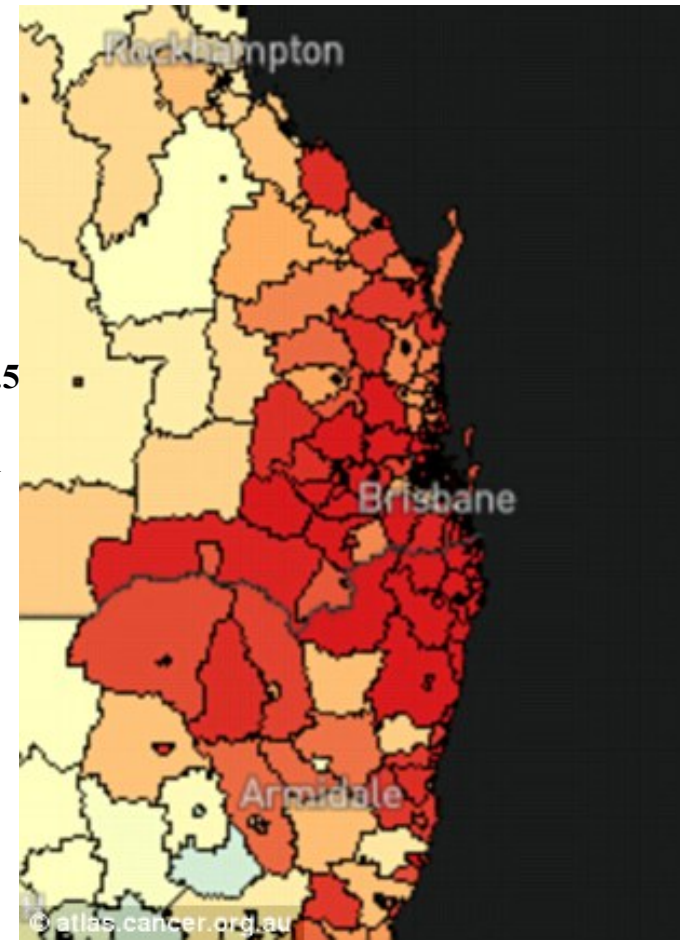


Figure 17: Skin cancer hotspots (Johnson 2018)

Major risk factors for skin cancer & ozone recovery initiative

High UV exposure is responsible for over 95% of the skin cancer. Depletion of the ozone layer further increase the UV exposure. Ozone layer was impacted due to the release of ozone-depleting substances (ODS) such as chlorofluorocarbons (CFCs). Australia is specifically impact due to its geographic location under the Southern Hemisphere's ozone hole. The ozone hole first identified in the 1980s, which is periodically forms over Antarctica, extends its effects into Australia. This elevated UV exposure which led to the highest skin cancer rates in the world.

To phase out ozone-depleting substances country has taken many strategies. As evident, Australia is one of the first countries to ratify the Montreal Protocol in 1989 (Australian Government Department of Climate Change, Energy, the Environment and Water 2024). The country also led by banning importation and manufacture of CFCs (Department of the Environment and Heritage 2001). Montreal Protocol was able to banned CFCs release to atmosphere globally in 2010 (Western et al. 2023) and their strategies leading to notable recovery of the ozone layer in the upper stratosphere. However, expect to recover the ozone layer even to 1980 level by year 2060 to 2066 over the Antarctic (UN environment programme 2023)

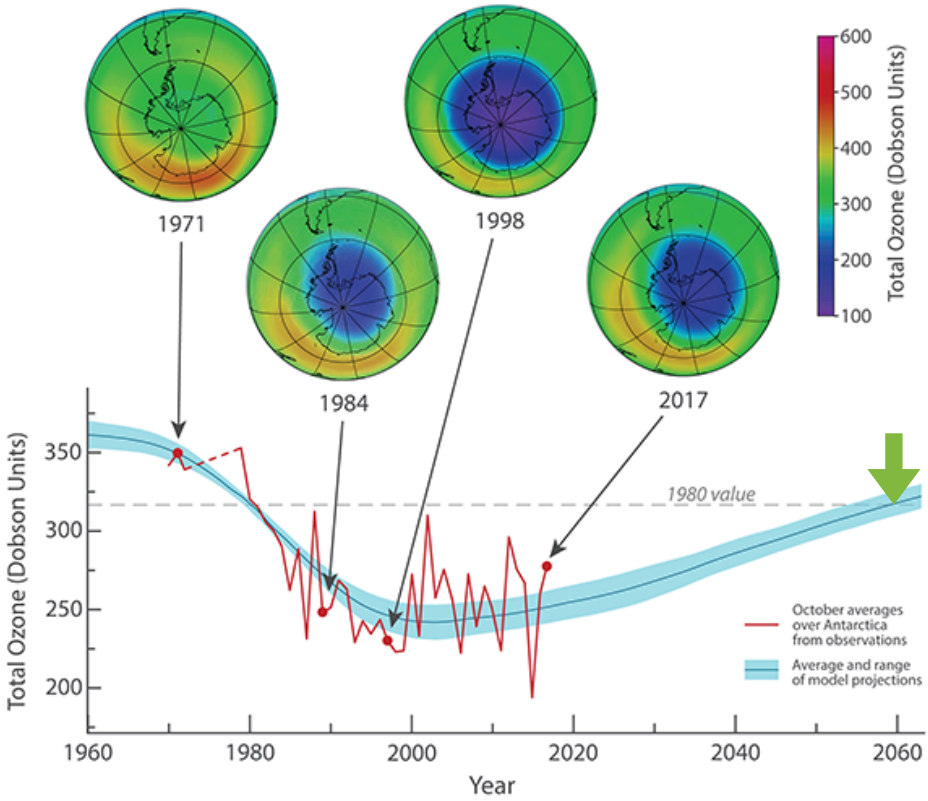


Figure 18: Ozon recovery pathway (NOAA Chemical Sciences Laboratory (CSL) 2020)

Other Key risk factors

HAVING EASILY BURN SKIN

OUTDOOR LIFE STYLES

FRECKLED SKIN/ MULTIPLE MOLES

(The Australasian College of Dermatologists 2022)

Predominant of the fair skin population is also one of the major reason for high skin cancer rates in Australia (Cancer Council 2023). According to figure 4 fair skin which burns easily and has higher susceptibility for skin cancer. Risk is high among skin types 1 to 3 as these skin types could damage easily from UV radiation due to less melanin. Dark skin depicts low risk. However lack of vitamin D will increase the risk.

Vast landscapes and mesmerizing coastlines of Australia attract for out door lifestyle such as surfing, barbecues, various sports, and beach gatherings. It will expose individuals to UV rays long hours which will increase the risk.

In addition to that freckled skin or skin consists with multiple moles has high risk of developing skin cancer.

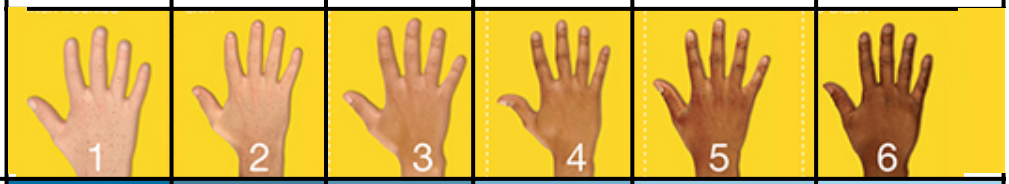
Skin Colour	Very fair Pale white Often freckled	Fair white skin	Light brown	Moderate brown	Oark brown	Deeply pigmented dark brown to black
						
UV sensitivity & tendency to burn	"Highly sensitive" "Always burns, never tans"	"Very sensitive" "Burns easily, tans minimally"	"Sensitive" "Burns moderately, usually tans"	"Less sensitive" "Burns minimally, tans well"	"Minimal sensitivity" "Rarely burns"	"Minimal sensitivity" "Never burns"
Skin cancer risk	"Greatest risk of skin cancer"	"High risk of skin cancer"	"High risk of skin cancer"	"At risk of skin cancer"	"Skin cancers are relatively rare. But those that occur are often detect later, more dangerous stage. Increased risk of low vitamin D levels"	"Skin cancers are relatively rare. But those that occur are often detect later, more dangerous stage. Increased risk of low vitamin D levels"

Figure 19: Skin cancer effect on different layers of the skin (Hasan et al. 2023)

Types of skin cancer

Skin cancer is primarily categorized to two types Non-Melanoma (NMSC) and Melanoma skin cancer (Gruber, Shah & Zito 2020). NMSC are the most common skin cancer in the country, consists Basal Cell Carcinoma (BCC) which accounts 70% and Squamous Cell Carcinoma (SCC) accounts 30% (Cancer Council 2024).

NMSC accounts 99% of all skin cancers (Ciążyńska et al. 2021) with over 400 000 new cases report annually (Australian Institute of Health and Welfare 2024). It is expected to rise continuously by 2%-6% annually. BCC shown high five year survival rate of more than 99%. Early stages (Stage 0 and Stage 1)of SCC depicts a higher survival rate 95%-98% and drop to 34-46% in metastatic stage (stage IV).

NMSC treatment cost reached AU$824 million in 2015. Furthermore, GP consultations also increased which highlight the growing burden in healthcare.

Melanoma accounts 1–2% of all skin cancers (Cancer Council NSW 2021). However, it is the most aggressive type (Better Health Channel 2014) responsible for 75% of skin cancer death (Healthdirect Australia 2019 ; (The Australasian College of Dermatologists 2022). As evident one person will die every 5 hours due to skin cancer (The Australasian College of Dermatologists 2022). Melanoma is more common in age 20-39 years (Melanoma Institute Australia 2023). Figure 4 depicts how it impacts on the inner skin layers and it can be impact on wider layer of the skin and also organs. 5 year survival rates will drop 97% in early stages (1 & 2) to 30% in stage IV. Anticipated annual cost of treatment was AU$272 million in 2017 (Elliott et al. 2017). This highlight necessitating prevention strategies and early diagnosing tools.

Figure 20: Skin cancer types appearance on the skin (C. Kavitha et al. 2024)

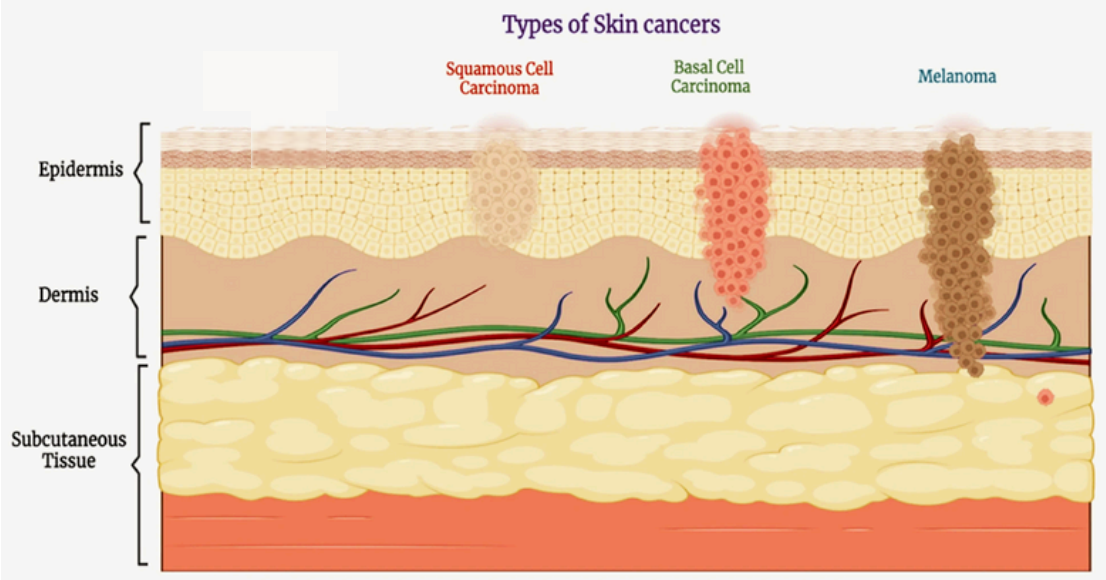
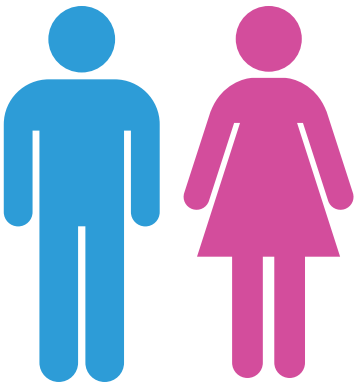


Figure 21: Skin cancer effect on different layers of the skin (Hasan et al. 2023)

Appendix 5



Melanoma is most aggressive skin cancer can spread anywhere in the body (Cancer Research UK 2020) and even able to damage internal organs. Considering gender it is identified that men are more likely to have melanoma skin cancer than women (Anthony.Augustine 2023). Similarly below chart also evident higher number of cases in men and also shows that risk of having melanoma will rise with the age.

It shows higher mortality rate in men, with a 1 in 84 chance of death. For women shows 1 in 284 mortality rate.

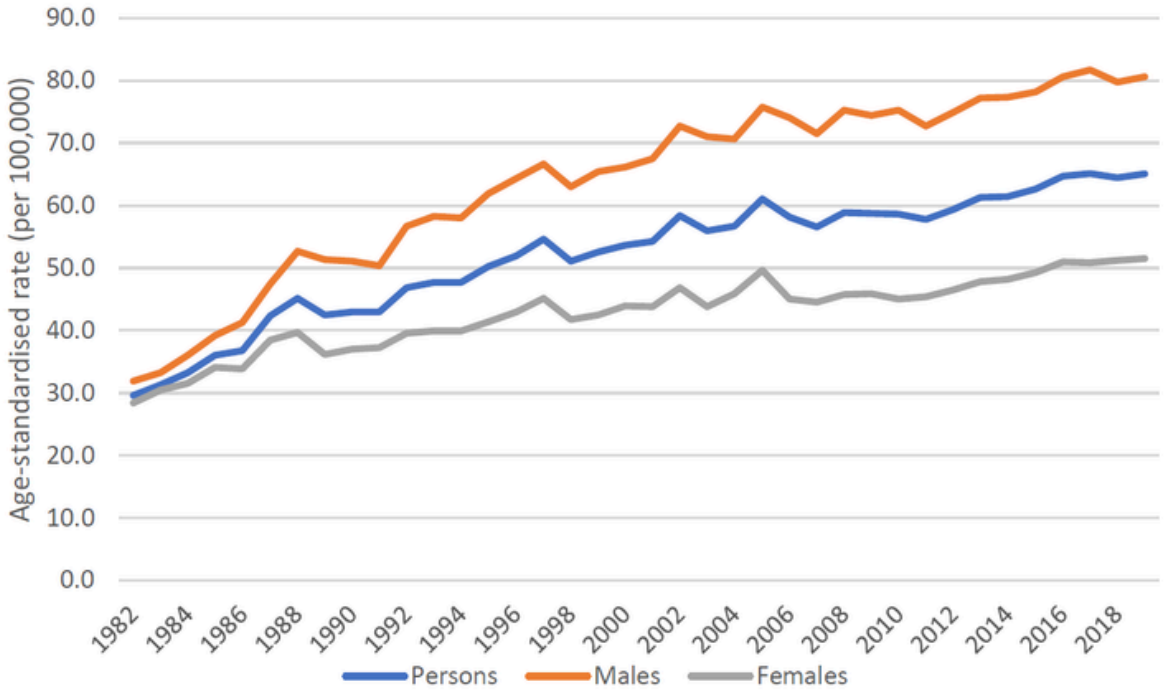


Figure 22: Melanoma skin cancer rates with age and gender (Cancer Australia 2022)

Data Sources

Author collected data from following platforms to use large data set of skin cancer lesion images to ensure diverse representation of three types of skin cancers.

- Kaggle- Skin cancer data set
  https://www.kaggle.com/datasets/mahdavi1202/skin-cancer
- Kaggle -Skin Disease Classification
  https://www.kaggle.com/datasets/riyaelizashaju/skin-disease-classification-image-dataset
- Roboflow- Skin-Disease-Four Dataset
  https://universe.roboflow.com/lums-szkgm/skin-disease-four/dataset/1
- Stanford AIMI - Skin disease dataset with diverse skin tone
  https://universe.roboflow.com/skripsi-1h4ty/skin-diseases-v2/dataset/1
- The University of Waterloo- Skin Image Data Set 1.zip
  https://uwaterloo.ca/vision-image-processing-lab/research-demos/skin-cancer-detection

The above are the original data sets used. But Author has done data pre-processing **(Appendix 7)** by transferred original data to clean and balance data set before analysis. Clean data set can be found in GitHub.
https://github.com/Hansani123/Data-6000-Skin-Cancer-

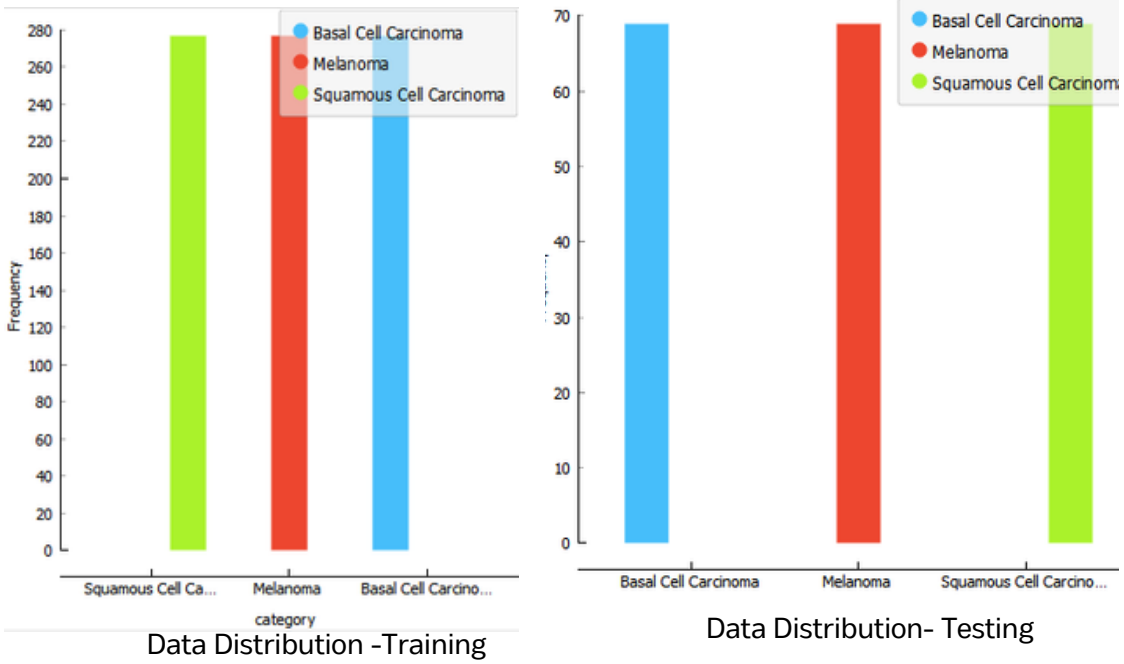
Data pre-processing

The below table demonstrate the number of skin cancer data collected by the author and how data is balanced and split to training and testing.

Data Source Site	Skin Cancer Disease			Balance data set preparation			Notes
	Melanoma (MEL)	Squamous Cell Carcinoma (SCC)	Basal Cell Carcinoma (BCC)	Melanoma (MEL)	Squamous Cell Carcinoma (SCC)	Basal Cell Carcinoma (BCC)	
Kaggle data set 1	52	192	845	52	192	148	
Kaggle data set 2	100	100	-	81	100	-	
Roboflow data set	437	-	376	82	0	149	
Stanford AIMI data set	20	54	49	20	54	49	2
The University of Waterloo	111	-	-	111	-	-	2
Total number of images	**720**	**346**	**1270**	**346**	**346**	**346**	
Note: 1	**346 is the minimum number of image data in skin cancer data types**						
Data split to training and testing							
Training 80% = 346*80%	277	277	277				
Testing 20% = 346*20%	69	69	69				

Figure 23 : Calculation of data balancing, and splitting into training and testing

Note 2 : After delete the duplicated images author used all the skin cancer image data in Stanford AIMI data set. Also used all the melanoma skin cancer images in the university of Waterloo data set since those two data sets are from more reliable sites. Only melanoma skin cancer data was available in the University of Waterloo data set. I selected the images randomly from Kaggle and Roboflow data sets in equal proportion. Then author combined the data by aggregating images of different skin cancer types and merged all the data sets into a unified collection for each skin cancer. Below graphs shows the balance training and testing data after pre-processing the data.



Data Distribution -Training

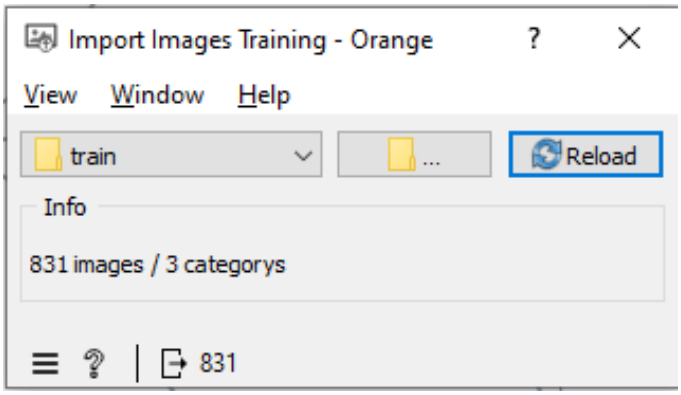
Data Distribution- Testing

Figure 24: Visualization of the balance data (Orange data mining)

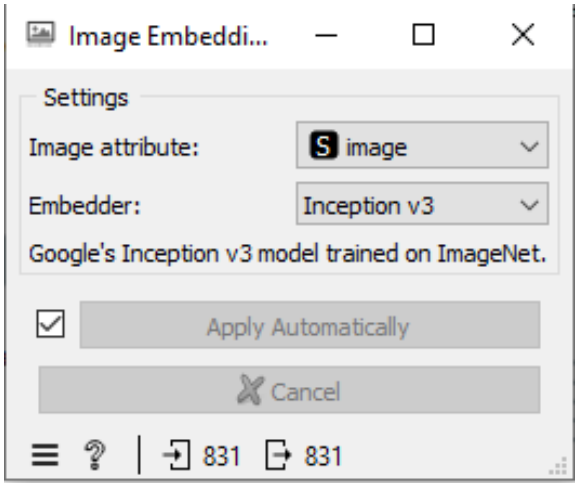
Appendix 9

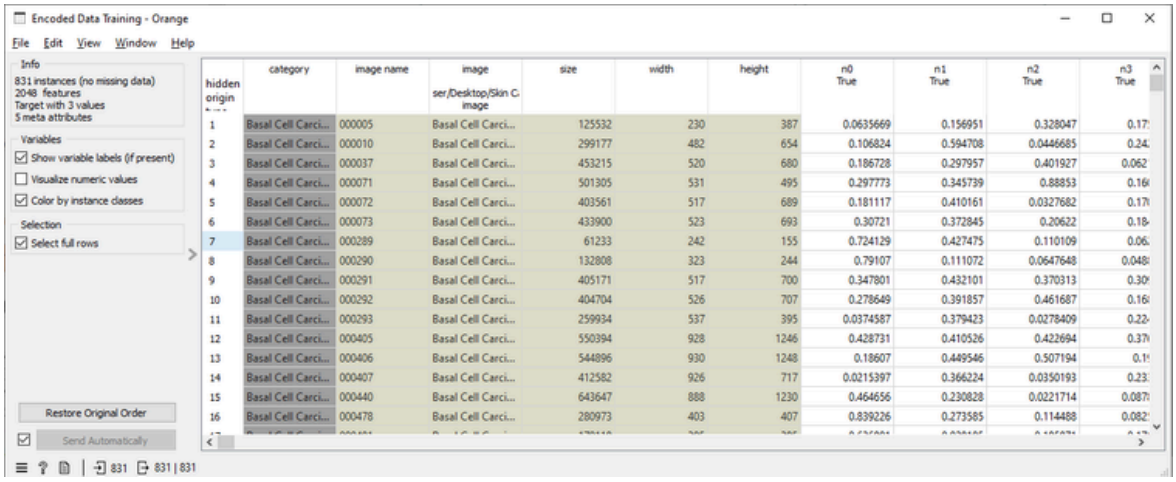
Data Analytics Methodology

Data import to the training model



Author used image embedding to transform images in vector representation. Then machine learning algorithm can read it using inception V3.

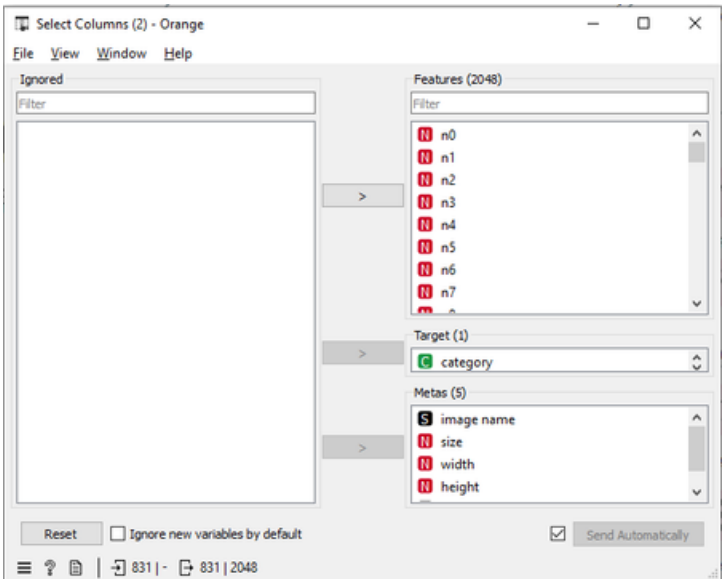




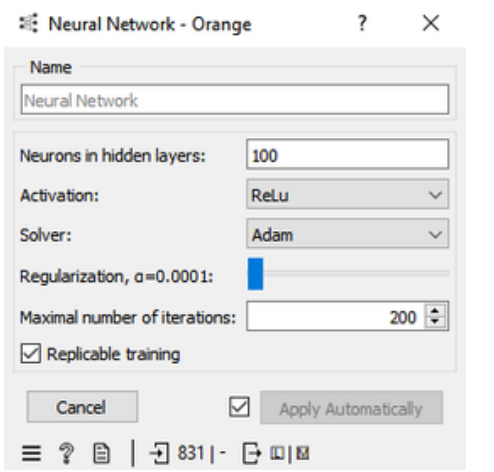
The images are easily understandable for humans but not for machines. Therefore, need to undergo the embedding process. In this process images transforms to vectors numbers which can be read by machine learning algorithms. Each number holds one property of image. Author used inception V3 to embed the image dataset, since it is deep neural network architecture designed for image recognition. Image embedding was applied to both training and testing data.

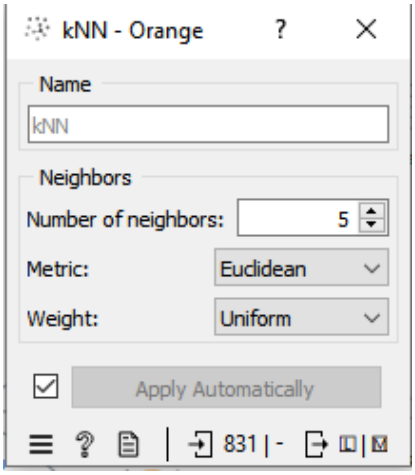
Appendix 09

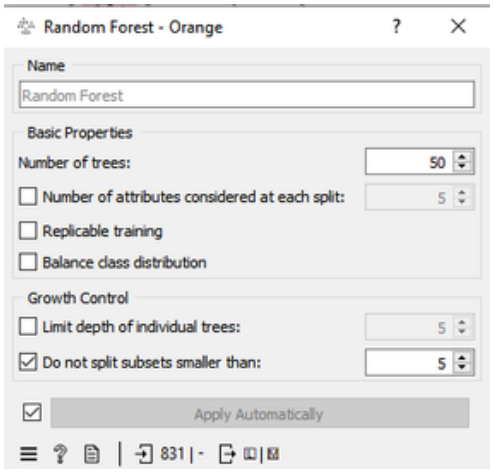
Author selected the target variable in the widget.

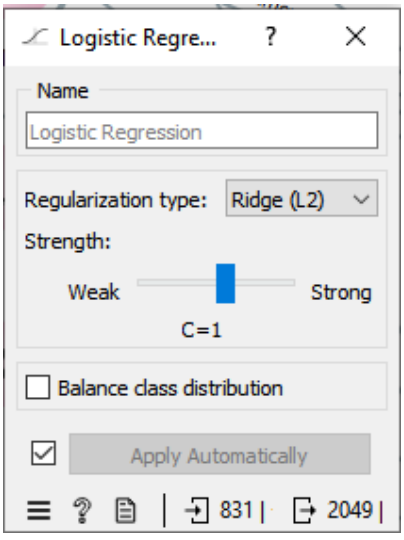


Models setup (Hyperparameter tuning)









Predictive analytics

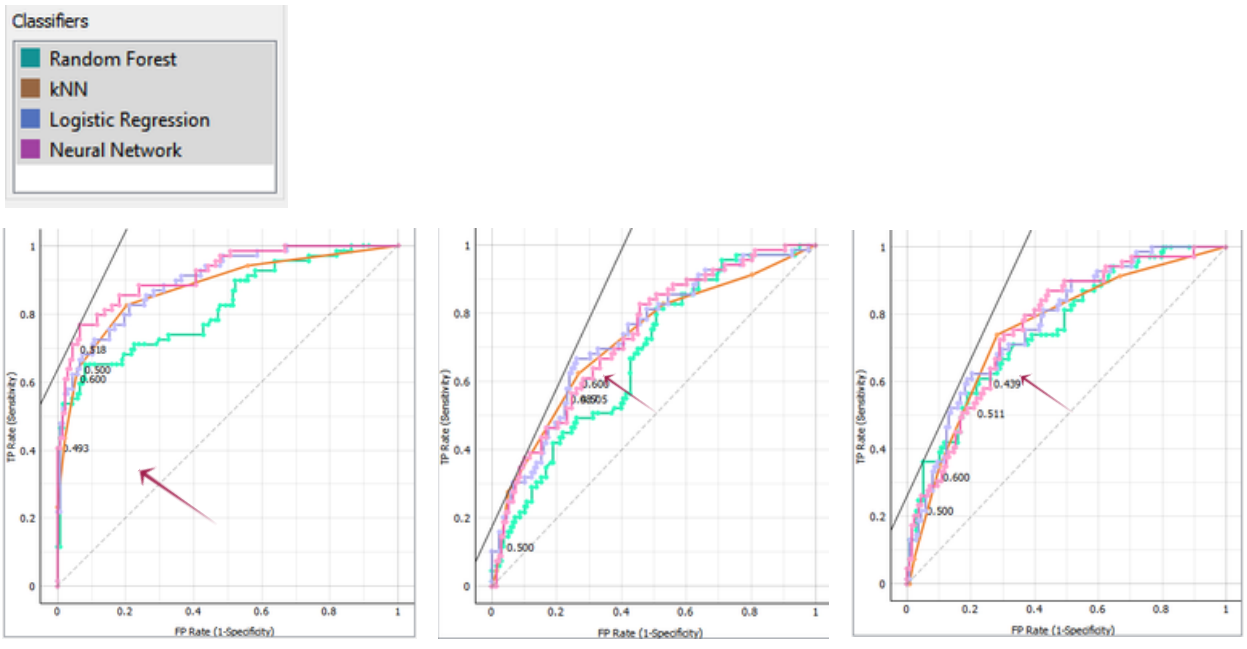
Testing data

It is important to assess how well models perform on test data which kept separately. It helps to obtain idea how models will perform with new data in real-world skin cancer classification scenarios in assisting doctors.

Figure 25 and 26 also Neural Networks shows the best results. Testing data shows slightly fewer performance metrics than training data.

Model	AUC	CA	F1	Prec	Recall	MCC
Random Forest	0.748	0.594	0.599	0.614	0.594	0.394
kNN	0.782	0.580	0.575	0.619	0.580	0.387
Logistic Regression	0.799	0.623	0.628	0.638	0.623	0.436
Neutral Network	0.803	0.618	0.624	0.634	0.618	0.429

Figure 25: Performance metrics for Testing data)



Melanoma (MEL)

Basal Cell Carcinoma (BCC)

Squamous Cell Carcinoma (SCC)

Figure 26: ROC curve analysis for testing data

Testing data

Figure 27 and 28 shows confusion matrix for the Random Forest and Neural Network models for testing data. NN model miss classified 79 cases out of 207 and Random forest model have even higher misclassified cases (88 cases)

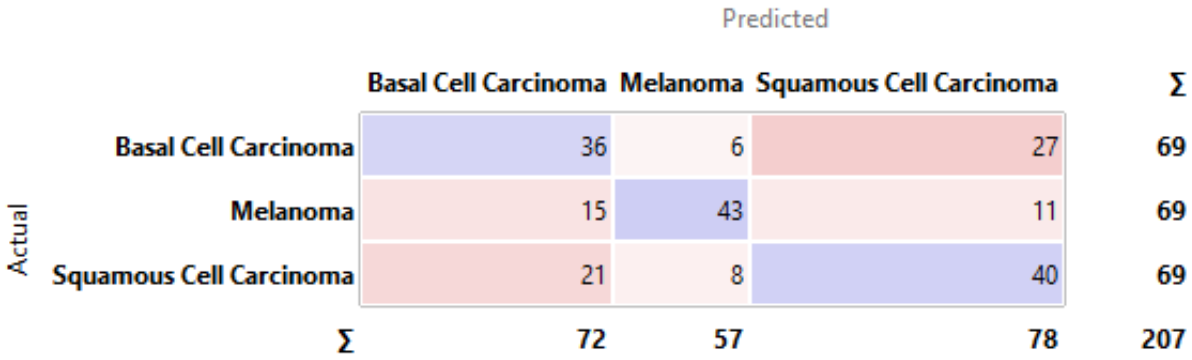


Figure 27: Confusion Matrix - Random forest (Testing dataset)

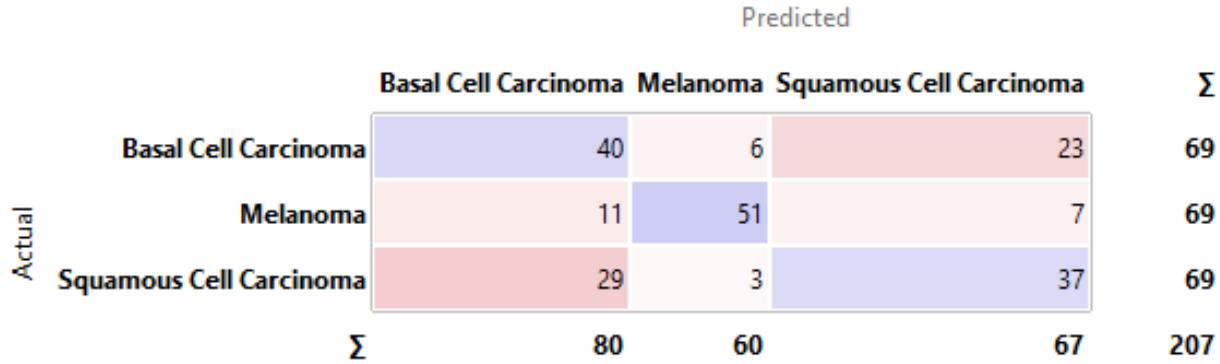


Figure 28: Confusion Matrix - Neural Network (Testing dataset)