DATA 6000
Capstone Project

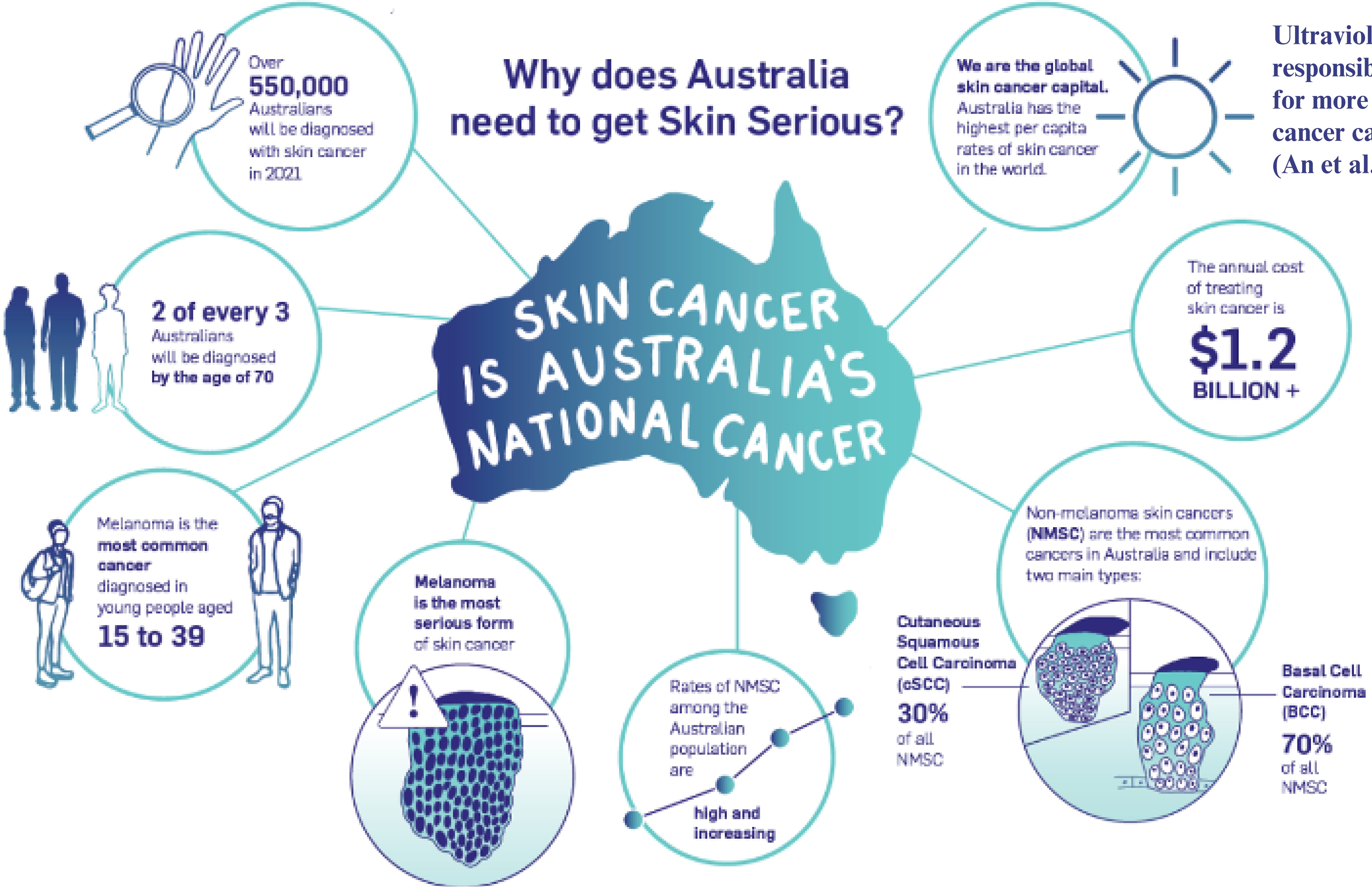
Healthcare Industry
Dermatology sector

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

(Source :EaTemp 2023)

PREPARED BY: HANSANI SEVANTHI
STUDENT ID: 1816338

Importance of addressing skin cancer



Ultraviolet radiation is responsible for more than 80% of skin cancer cases (An et al. 2021).

Figure 1: Why skin cancer is significant issue in Australia (mscan 2022)

Dermatalogy workforce distribution

Indicates a severe shortage of dermatologists in the country

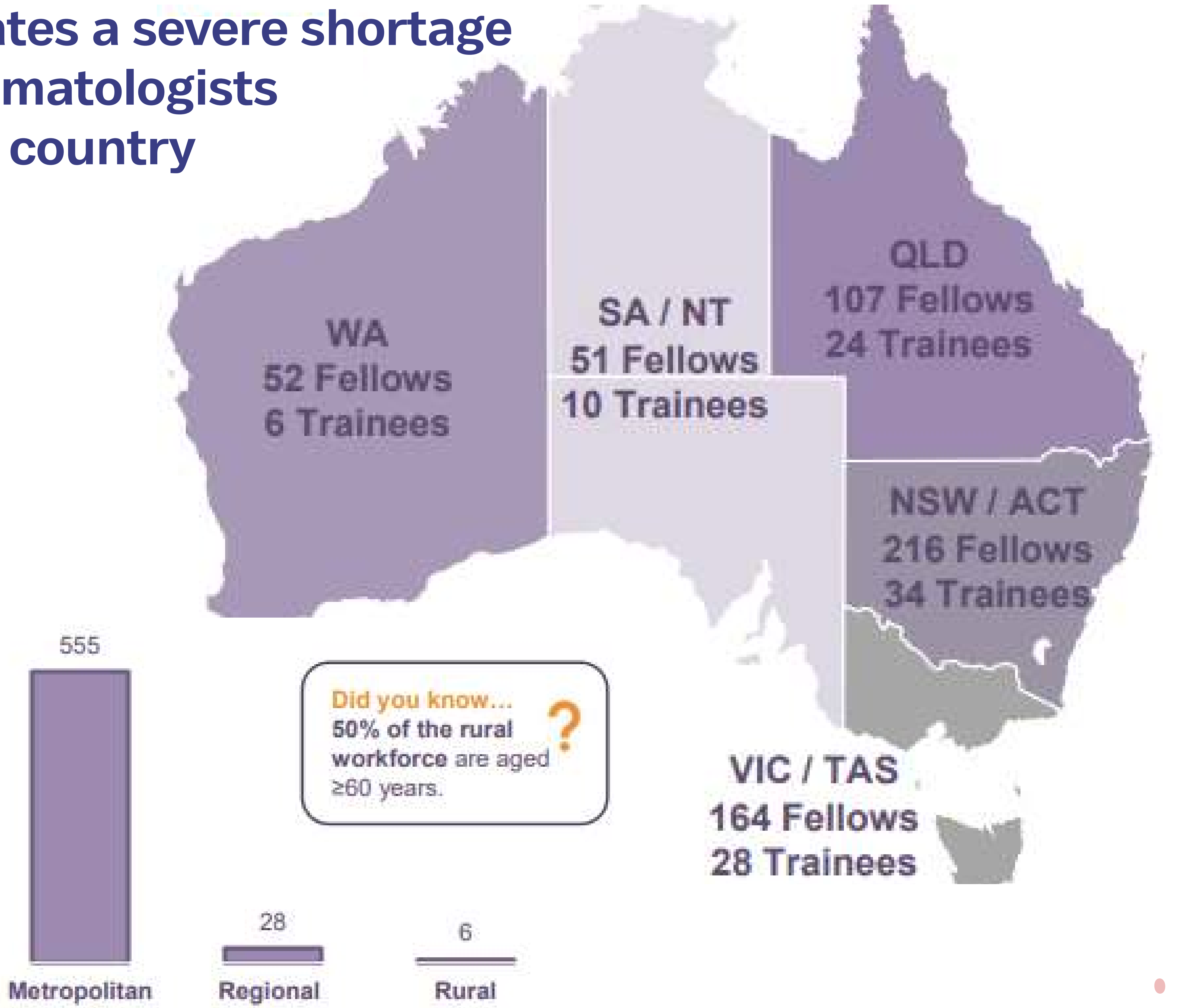


Figure 2: Dermatology Workforce Distribution in 2021 (The Australasian College of Dermatologists 2022)

Do You Know?

- "Australia has only 590 dermatologists serving a population of 26 million"(Sheppeard 2022).
- "Roughly 2 dermatologists for 100,000 Australians" (HIF Australia 2023).

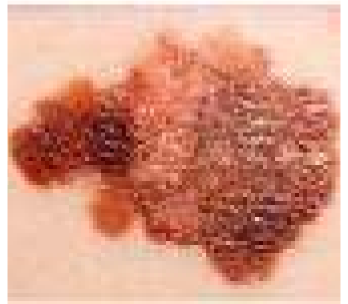

Source :(SLIDESGO 2024)

Objective

Predict skin cancer diseases of the patients visit GPs with trained AI and machine learning using image data set to address the severe shortage of the dermatologists in the dermatology sector in the healthcare industry

Types of skin cancer

Melanoma (MEL)



Basal Cell Carcinoma (BCC)

Squamous-cell skin cancer (SSC)

Figure 3: Skin cancer types external appearance on the skin (C. Kavitha et al. 2024)

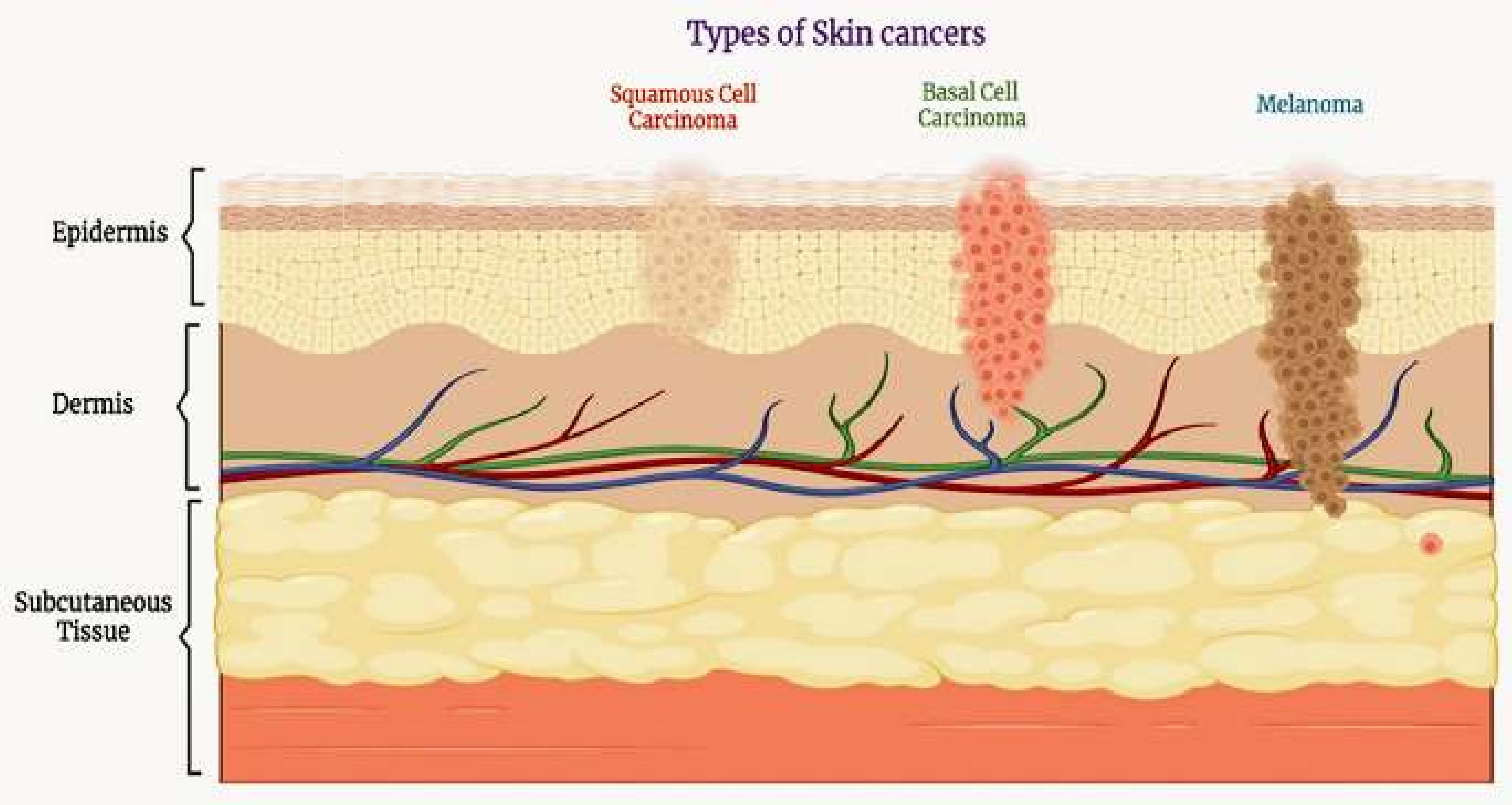


Figure 4: Skin cancer effect on different layers of the skin (Hasan et al. 2023)

Data Sources

- 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

Melanoma (MEL)

Basal Cell Carcinoma (BCC)

Squamous Cell Carcinoma (SCC)

Source :(SLIDESGO 2024)

Figure 5: Image Data Set

Work flow

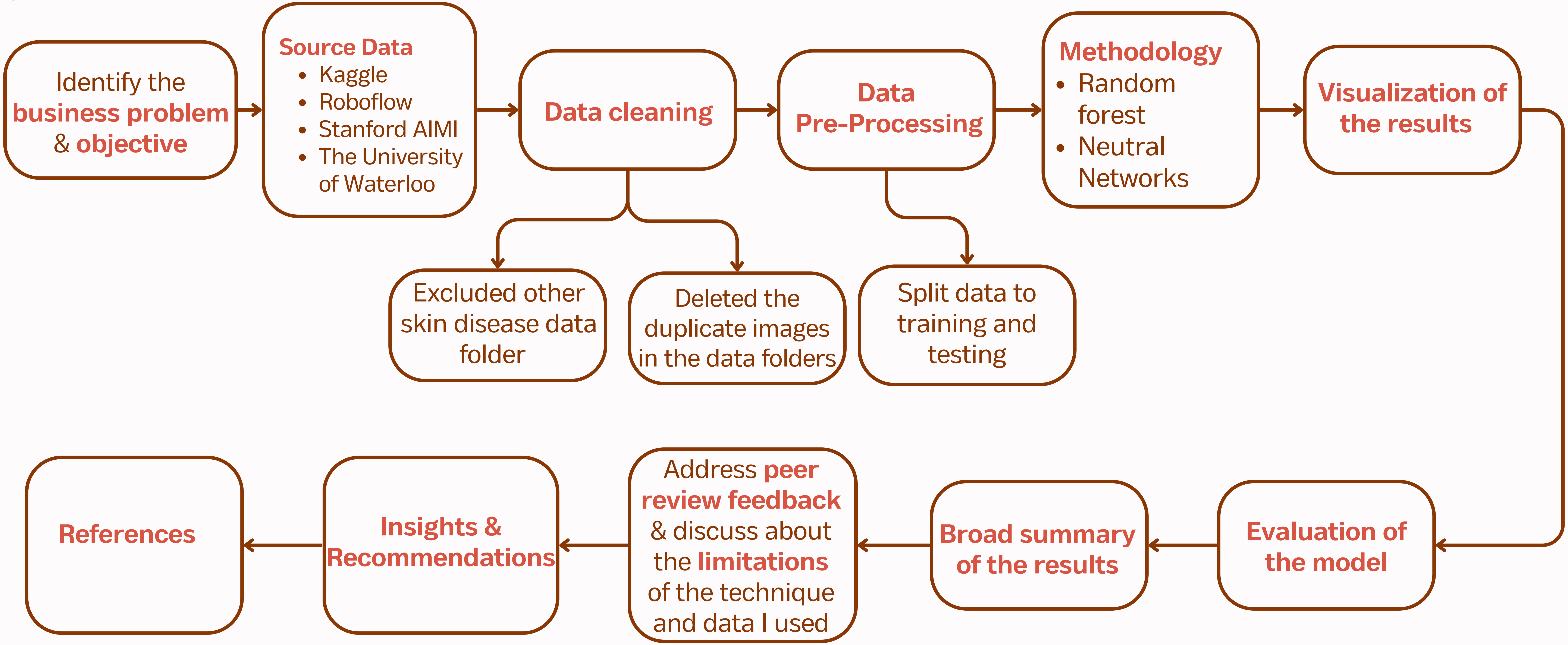


Figure 6: Workflow

Orange Data Mining

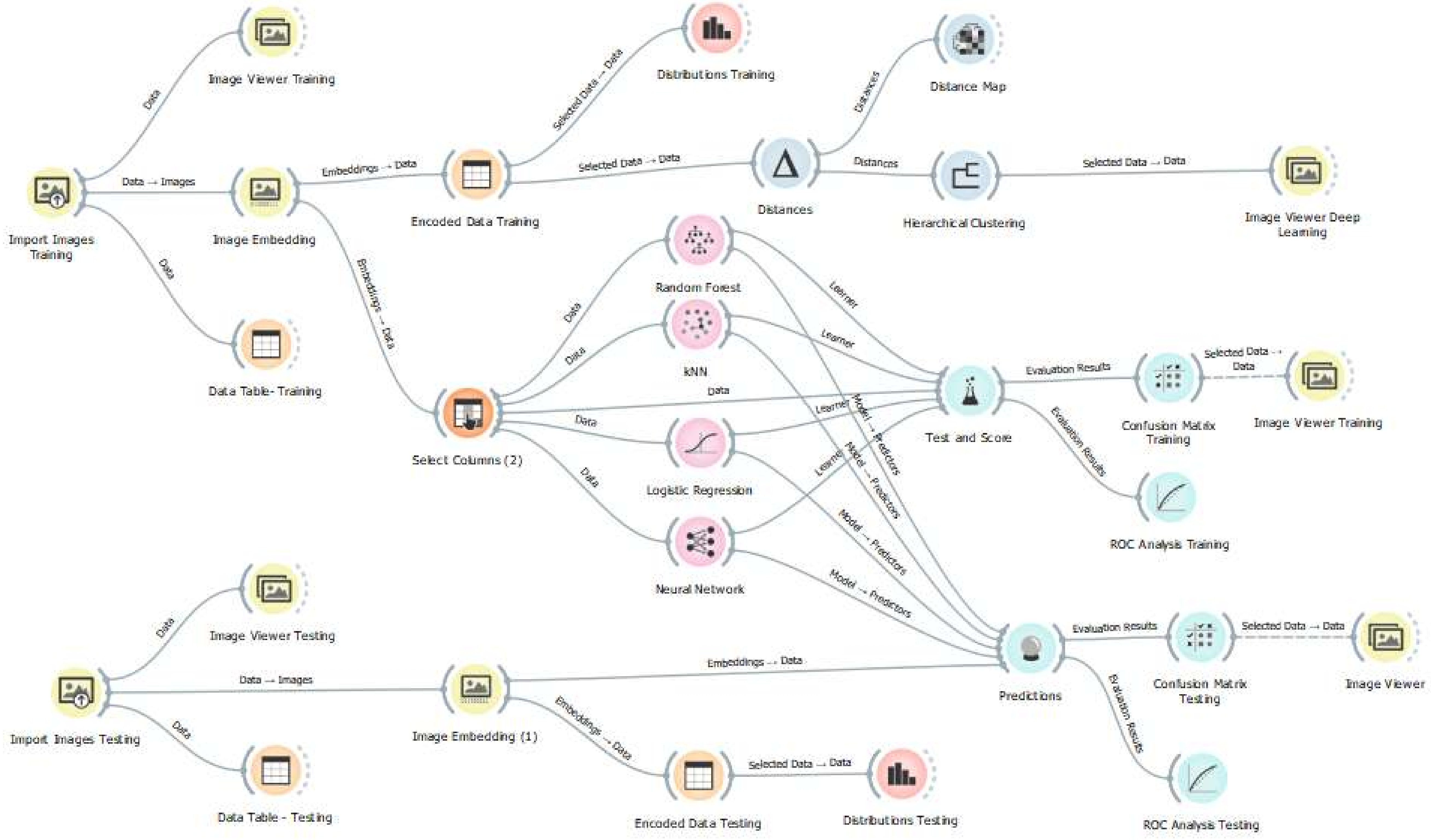
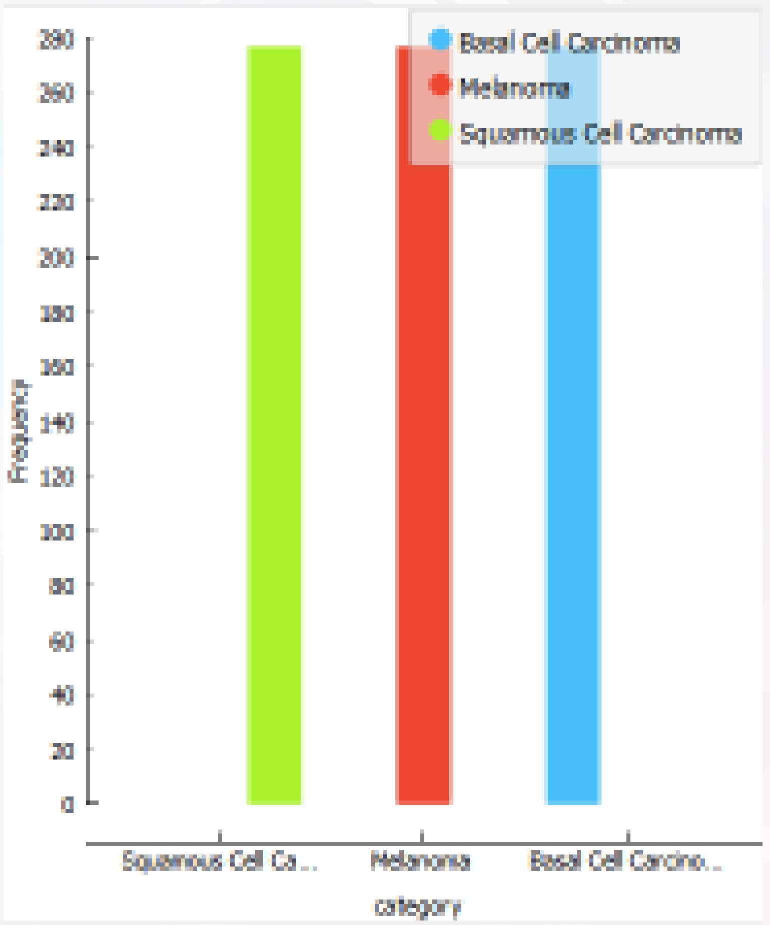


Figure 7: Orange Workflow (Orange data mining)

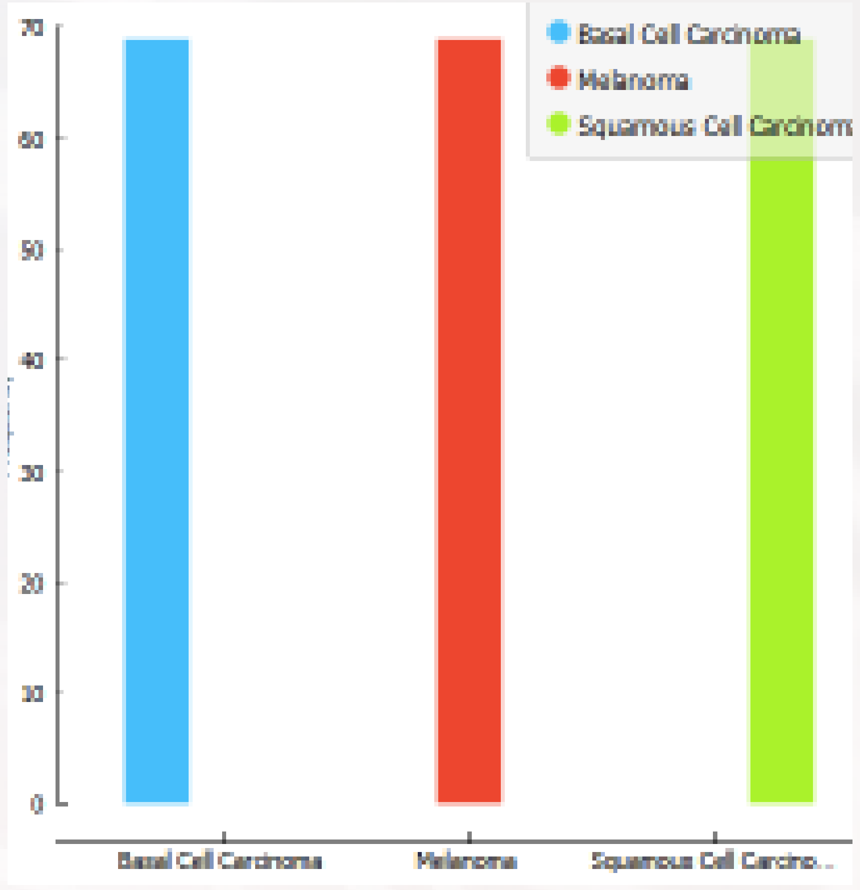
Data Pre-Processing

Data Source Site	Skin Cancer Disease		
	Melanoma (MEL)	Squamous Cell Carcinoma (SCC)	Basal Cell Carcinoma (BCC)
Kaggle data set 1	52	192	845
Kaggle data set 2	100	100	-
Roboflow data set	437	-	376
Stanford AIMI data set	20	54	49
The University of Waterloo	111	-	-
Total number of images	**720**	**346**	**1270**
Training 80% = 346*80%	**277**	**277**	**277**
Testing 20% = 346*20%	**69**	**69**	**69**

Figure 8: Calculation of data balancing, and splitting into training and testing



Data Distribution -Training



Data Distribution- Testing

Figure 9: Visualization of the balance data

Total Data
- 1038

Training Data
- 831

Testing Data
- 207

Descriptive Analytics

Hierarchical clustering - Dendrograms

- Hierarchical clustering helps to explore and summarize patterns within the data set by grouping different images based on the similarities on image features.
- It provides insights to the structure and relationships in the data set.

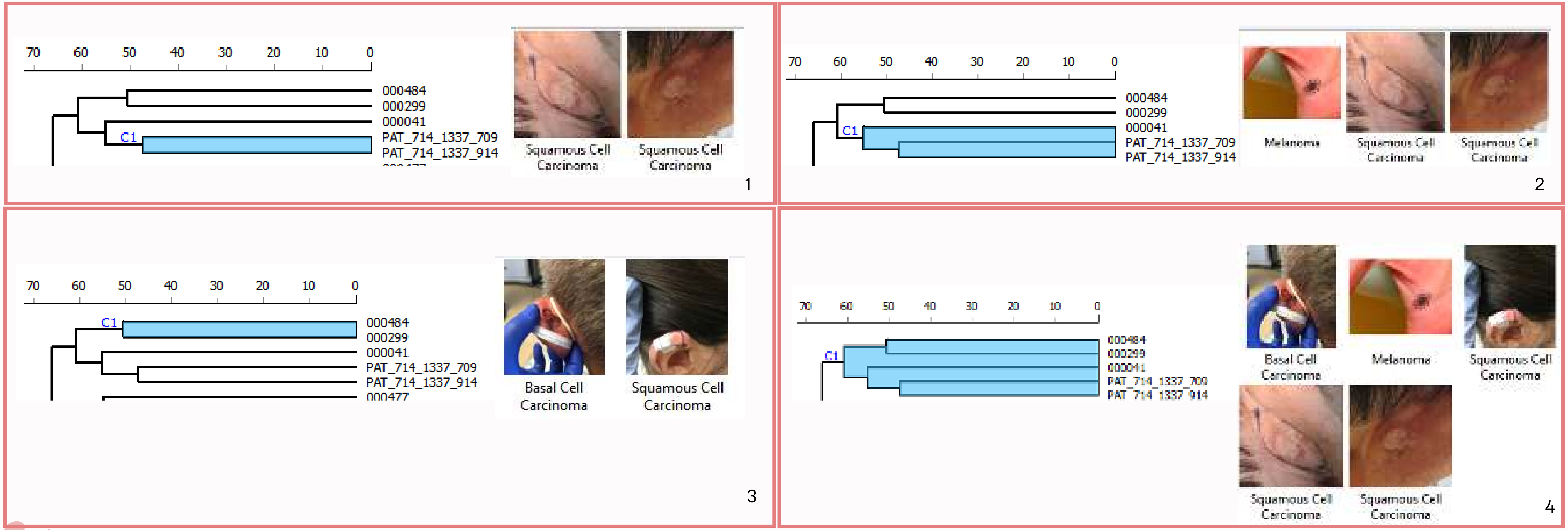


Figure 10: Visualization of hierarchical clustering of skin cancer image data

Predictive Analytics

For Random Forest & Neural Networks

Random Forest

Represent the correctly classified images

Actual \ Predicted	Basal Cell Carcinoma	Melanoma	Squamous Cell Carcinoma	Σ
Basal Cell Carcinoma	156	28	93	277
Melanoma	27	210	40	277
Squamous Cell Carcinoma	79	42	156	277
Σ	262	280	289	831

Figure 11: Confusion Matrix - Random forest

Random forest has the ability to handle categorical data without transforming them to numerical formats such as one-hot coding (Au 2018). Increasing the number of trees will led to robust predictions (Huh 2021).

Results interpretation

- Basal Cell Carcinoma : **The model predicted 156 cases correctly** out of 277 actual BCC cancer cases. Misclassified 121 cases as melanoma and SCC cancer cases.
- Melanoma : **210 melanoma cases were correctly predicted**. Misclassified 67 cases as other skin cancer types.
- Squamous Cell Carcinoma : **156 SCC cases were correctly predicted.** 121 cases misclassified as other skin cancer types.

Neural Networks

Actual \ Predicted	Basal Cell Carcinoma	Melanoma	Squamous Cell Carcinoma	Σ
Basal Cell Carcinoma	172	20	85	277
Melanoma	23	232	22	277
Squamous Cell Carcinoma	71	29	177	277
Σ	266	281	284	831

Figure 12: Confusion Matrix - Neural networks

Results interpretation

- Basal Cell Carcinoma : **The model predicted 172 cases correctly**. Misclassified 105 cases as melanoma and SCC cancer cases.
- Melanoma : **232 melanoma cases were correctly predicted**. Misclassified 45 cases as other skin cancer types.
- Squamous Cell Carcinoma : **177 SCC cases were correctly predicted.** 100 cases misclassified as other skin cancer types.

Predictive Analytics

Neural Networks

Confusion matrix and correctly identified BCC, MEL and SCC cancer cases

Predicted

Actual	Basal Cell Carcinoma	Melanoma	Squamous Cell Carcinoma	Σ
Basal Cell Carcinoma	172	20	85	277
Melanoma	23	232	22	277
Squamous Cell Carcinoma	71	29	177	277
Σ	266	281	284	831

Figure 13: Correctly predicted BCC cases

Predicted

Actual	Basal Cell Carcinoma	Melanoma	Squamous Cell Carcinoma	Σ
Basal Cell Carcinoma	172	20	85	277
Melanoma	23	232	22	277
Squamous Cell Carcinoma	71	29	177	277
Σ	266	281	284	831

Figure 14: Correctly predicted melanoma cases

Predicted

	Basal Cell Carcinoma	Melanoma	Squamous Cell Carcinoma	Σ
Basal Cell Carcinoma	172	20	85	277
Melanoma	23	232	22	277
Squamous Cell Carcinoma	71	29	177	277
Σ	266	281	284	831

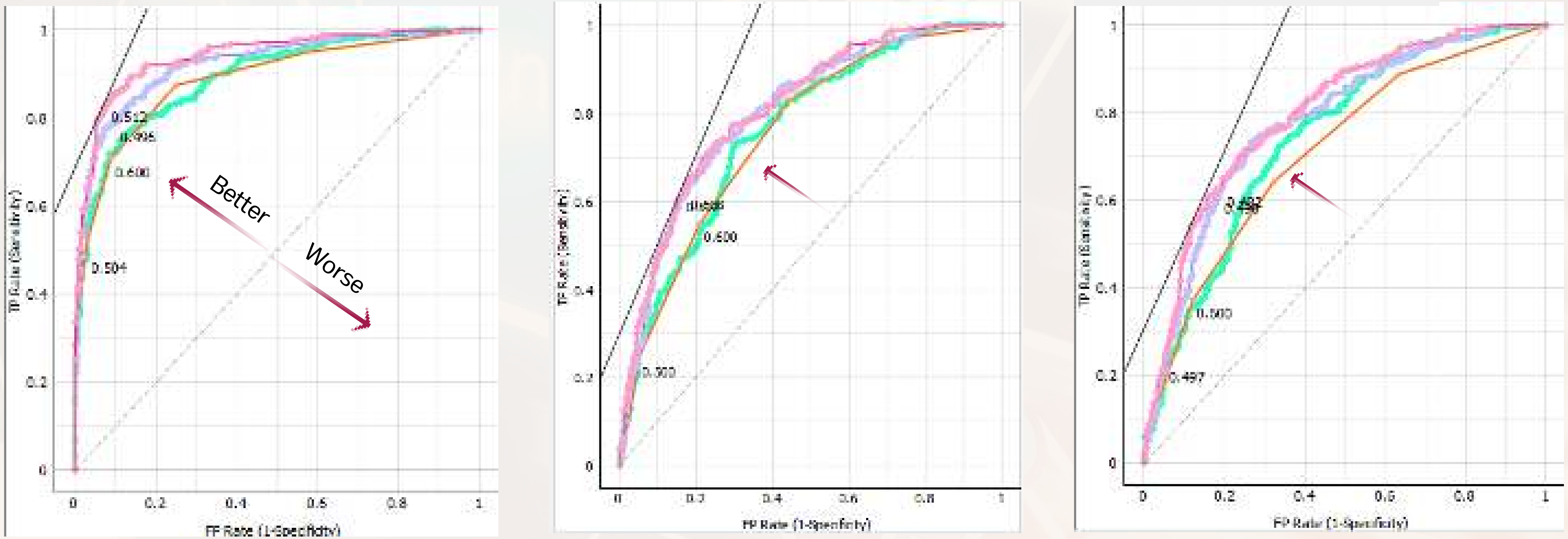
Figure 15: Correctly predicted SCC cases

Model Evaluation for Training Data

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 16: Performance metrics for predictive models for training data (orange data mining)

Classifiers
- Random Forest
- kNN
- Logistic Regression
- Neural Network



Melanoma (MEL)

Basal Cell Carcinoma (BCC)

Squamous Cell Carcinoma (SCC)

Figure 17: AUC- ROC curve analysis for the skin cancer data set for training data (orange data mining)

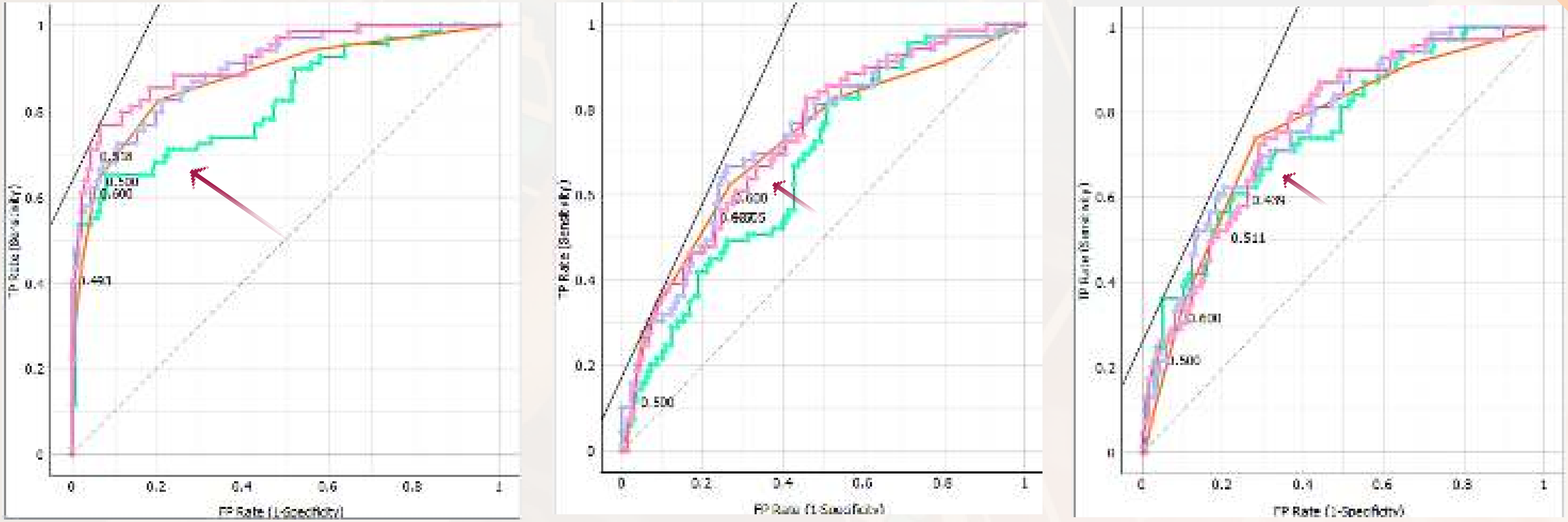
Model Evaluation for Testing Data

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 18: Performance metrics for predictive models for testing data (orange data mining)

Classifiers

- Random Forest
- kNN
- Logistic Regression
- Neural Network

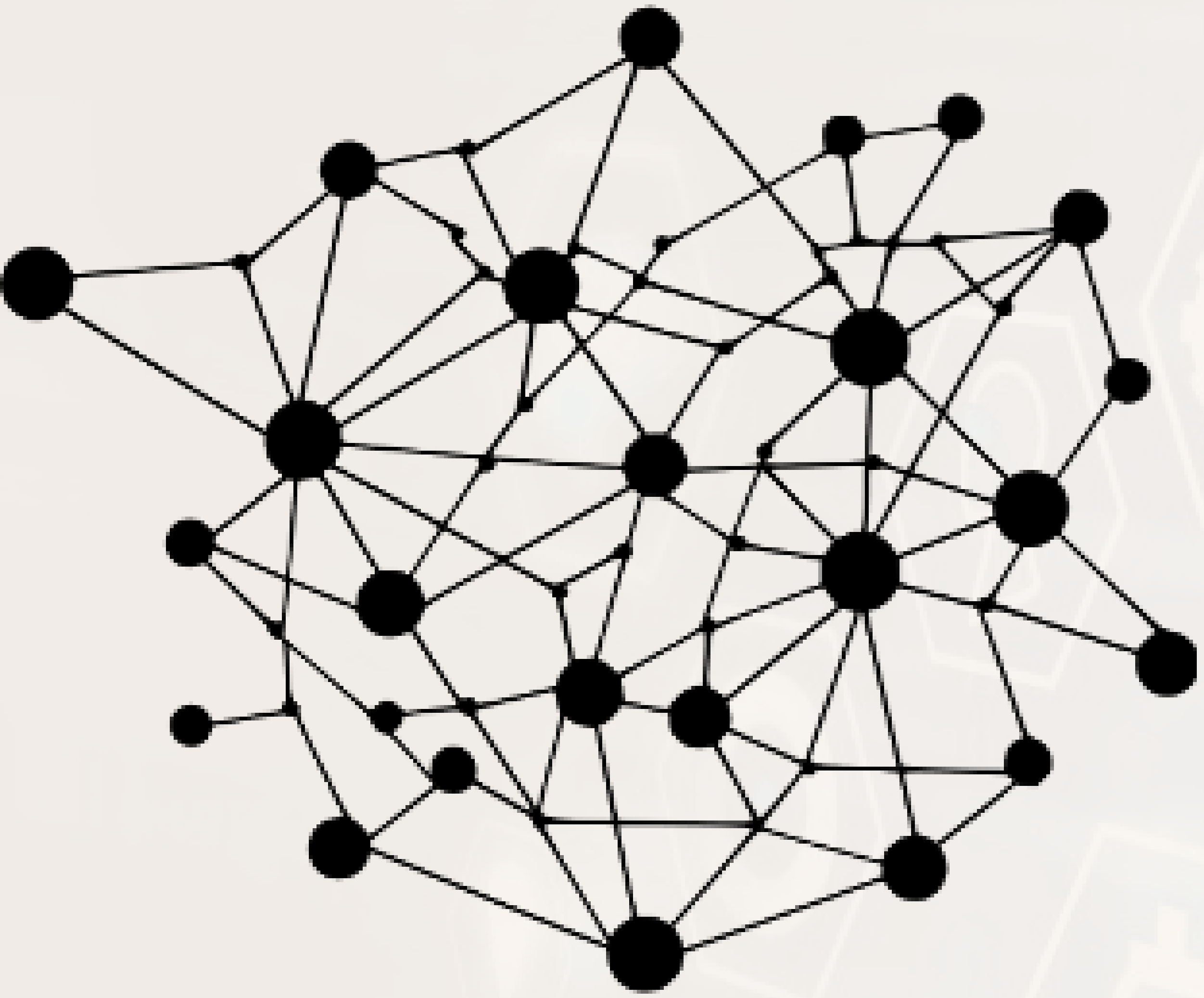


Melanoma (MEL)

Basal Cell Carcinoma (BCC)

Squamous Cell Carcinoma (SCC)

Figure 19: AUC- ROC curve analysis for the skin cancer data set for testing data (orange data mining)



Best ML Model

Neural Network

Accuracy = 84.9%

Precision = 69.8%
% of true predicted positives that are actually correct (Bhandari 2020).

Recall = 69.9%
% of true positive predictions among all the true positive cases (Wei 2024).

(Source :EaTemp 2024)

Peer Review Feedback & Limitations

(Source: FLATICON 2020)

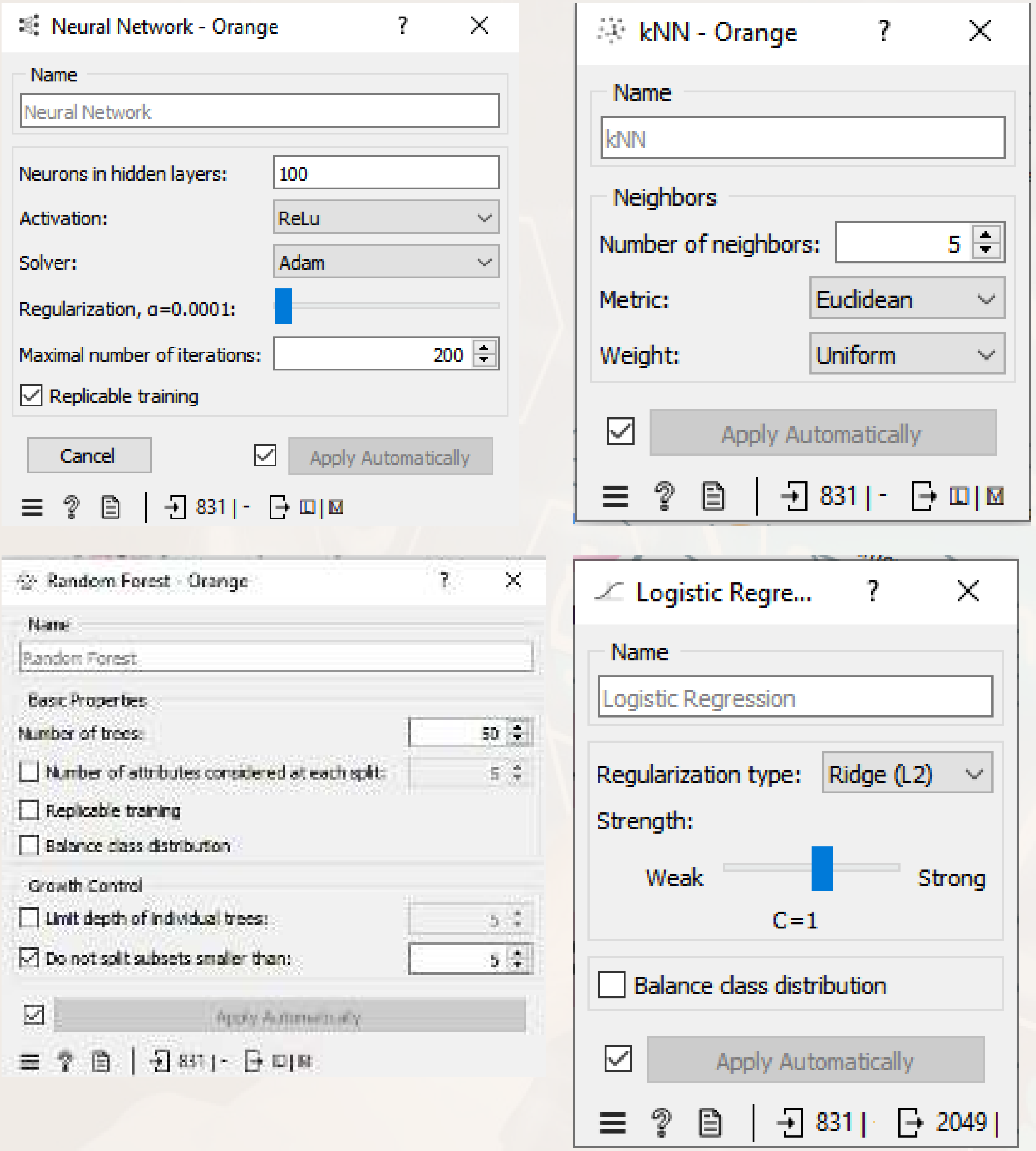


Figure 20: Hyper Parameter optimization (Arzola 2020)

Peer Review

- Pema's comments were very helpful. She suggested to tune hyperparameters using Grid search.
- However, in Orange software unable to use grid search for hyperparameter tuning.
- Optimized hyperparameters based on guidelines of a journal article (figure 20).

Limitations

- Neural Networks has lack of ability to capture spatial relationships effectively. For instance how the colors distributed across lesion, changes in the appearance of the skin surface.
- Convolutional Neural Networks CNN has the ability to capture spatial relationships. However, CNN can perform in programming language.
- Limited access to the data sets with all 6 types of skin tones. Only 2 data sets were able to find with different skin tones

Insights

- The models high accuracy rate (84.9%) and moderate number of positive out comes from the confusion matrix shows models accuracy in diagnosing skin cancer early to improve patients health out comes.
- Better precision rate (69.8%) and low false number from confusion matrix shows that this model reduce unnecessary medical procedures and improve patient care while reducing cost.
- Clinical professionals and dermatologists are able to prioritize high risk patients using models accuracy and recall rates.

(Source: FAMAGUSTA NEWS 2022)

Recommendations

- Either Neutral Network (NN) or CNN can be used for this image dataset to make predictions about skin cancer.
- Obtain image data from the data sets which provide the balance data set of the patients with different skin tones.
- Collaborate with clinical and medical professionals to validate the findings of the project.

References

- An, S, Kim, K, Moon, S, Ko, K-P, Kim, I, Lee, JE & Park, SK 2021, 'Indoor Tanning and the Risk of Overall and Early-Onset Melanoma and Non-Melanoma Skin Cancer: Systematic Review and Meta-Analysis', *Cancers*, vol. 13, no. 23, p. 5940, viewed 2 September 2024 <https://pubmed.ncbi.nlm.nih.gov/34885049/>.
- Arzola, A 2020, 'Modified Oregonator: an Approach from the Complex Networks Theory', *REVISTA MEXICANA DE INGENIERÍA BIOMÉDICA*, viewed 1 September 2024, <https://www.researchgate.net/publication/349199357_Modified_Oregonator_an_Approach_from_the_Complex_Networks_Theory>.
- Au, T 2018, 'Random Forests, Decision Trees, and Categorical Predictors: The "Absent Levels" Problem', *Journal of Machine Learning Research*, vol. 19, pp. 1–30, viewed 28 August 2024, <https://jmlr.csail.mit.edu/papers/volume19/16-474/16-474.pdf>.
- Bhandari, A 2020, *Confusion Matrix in Machine Learning*, Analytics Vidhya, viewed 27 August 2024, <https://www.analyticsvidhya.com/blog/2020/04/confusion-matrix-machine-learning/#:~:text=Precision%20tells%20us%20how%20many>.
- C. Kavitha, S. Priyanka, M. Praveen Kumar & Kusuma, V 2024, 'Skin Cancer Detection and Classification using Deep Learning Techniques', *Procedia Computer Science*, vol. 235, Elsevier BV, pp. 2793–2802, viewed 21 August 2024, <https://www.sciencedirect.com/science/article/pii/S1877050924009438?ref=pdf_download&fr=RR-2&rr=8bc585889de43056>.
- EaTemp 2024, *Dermatology PowerPoint Presentation Template & Google Slides Theme*, EaTemp - Free and Professional Resume and Presentation Templates, viewed 24 July 2024, <https://eatemp.com/dermatology-powerpoint-presentation-template/>.

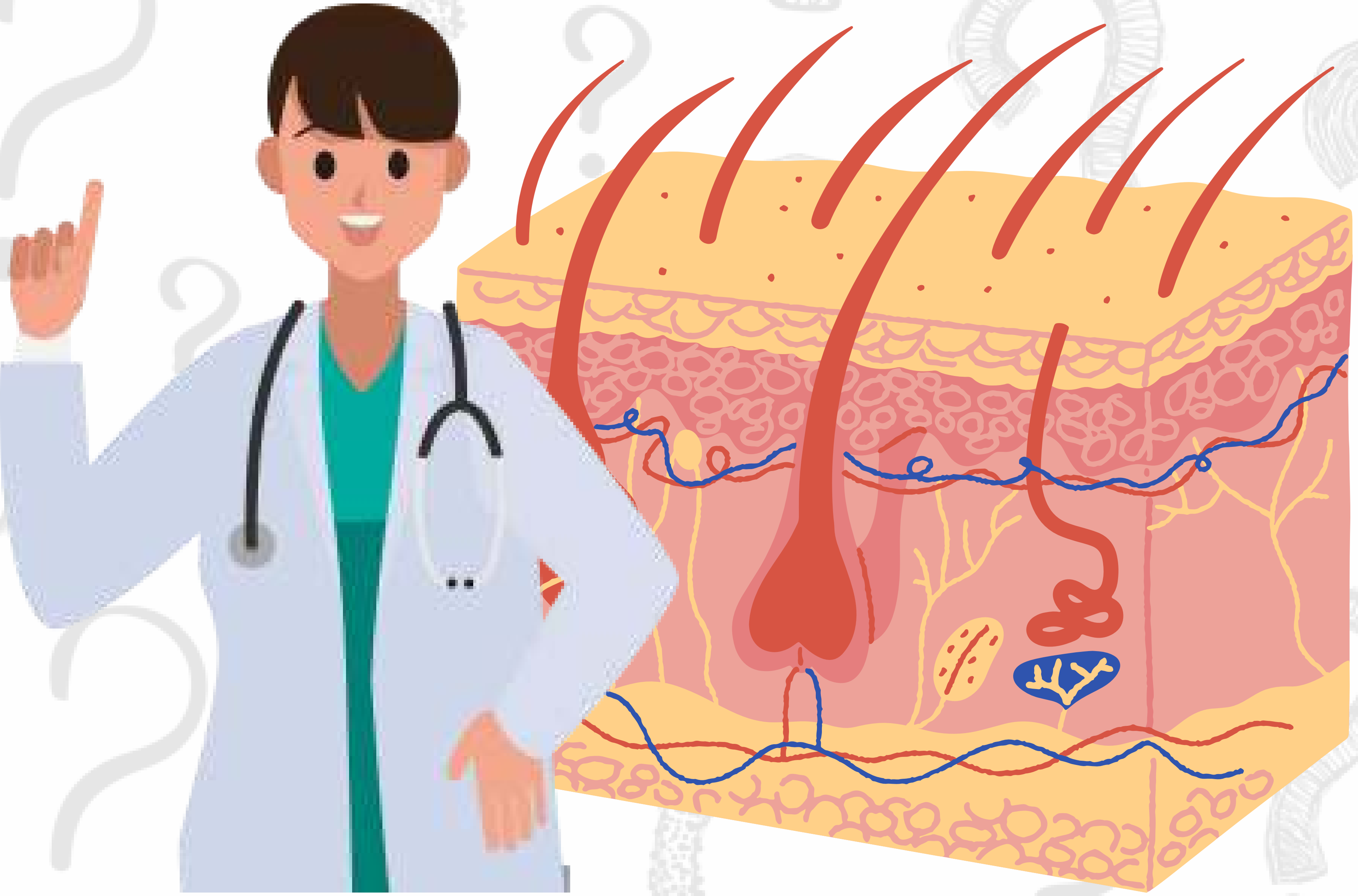
References

- FAMAGUSTA NEWS 2022, ‘*Pandemic personal doctors absent - They get caught and do not answer*’ - *Famagusta News, Famagusta News*, viewed 15 August 2024, <https://en.famagusta.news/news/apontes-oi-prosopikoi-giatroi-apo-pandimia-tous-pianoun-kai-den-apantoun>.
- FLATICON 2020, *Conversation Icon* - 2975684, Flaticon, <https://www.flaticon.com/free-icon/conversation_2975684?related_id=2975723&origin=search>.
- Hasan, N, Arif Nadaf, Imran, M, Umme Jiba, Sheikh, A, Almalki, WH, Salem Salman Almujri, Mohammed, Y, Prashant Kesharwani & Farhan Jalees Ahmad 2023, ‘Skin cancer: understanding the journey of transformation from conventional to advanced treatment approaches’, *Molecular Cancer*, vol. 22, BioMed Central, no. 1, viewed 4 September 2024, <https://molecular-cancer.biomedcentral.com/articles/10.1186/s12943-023-01854-3>.
- HIF Australia 2023, *Dermatology & Private Health Insurance in Australia - HIF*, Hif.com.au, viewed 21 August 2024, <https://healthhub.hif.com.au/general-medical/is-dermatology-covered-by-private-health-insurance-what-you-need-to-know>.
- Huh, K 2021, *Surviving In a Random Forest with Imbalanced Datasets*, SFU Professional Computer Science, viewed 22 August 2024, <https://medium.com/sfu-cspmp/surviving-in-a-random-forest-with-imbalanced-datasets-b98b963d52eb>.
- mscan 2022, *Facts and figures*, MSCAN, viewed 21 August 2024, <https://mscan.org.au/learning-hub/skin-cancer/facts-and-figures/>.

References

- Sheppeard, A 2022, *Australia's dermatology workforce crisis*, Dermatology Republic, viewed 21 August 2024, <https://www.dermatologyrepublic.com.au/australias-dermatology-workforce-crisis/544>.
- SLIDESGO 2024, *Skin Cancer Disease* | Google Slides and PowerPoint template, Slidesgo.com, viewed 11 August 2024, <https://slidesgo.com/theme/skin-cancer-disease>.
- The Australasian College of Dermatologists 2022, *The Australasian College of Dermatologists Federal Pre-Budget Submission 2022-23 Accessible and equitable dermatology care for all patients and communities, Australian Government The Treasury,* viewed 8 August 2024, <https://treasury.gov.au/sites/default/files/2022-03/258735_australasian_college_of_dermatologists.pdf>.
- The DermEngine Team 2018, *Top Ways AI Is Revolutionizing Dermatology*, Dermengine.com, viewed 2 September 2024, <https://www.dermengine.com/blog/artificial-intelligence-ai-dermoscopy>.
- Unsplash 2024, *Beautiful Free Images & Pictures, Unsplash, Unsplash*, viewed 24 July 2024, <https://unsplash.com/>.
- Wei, D 2024, *Essential Math for Machine Learning: Confusion Matrix, Accuracy, Precision, Recall, F1-Score*, Medium, viewed 27 August 2024, <https://medium.com/@weidagang/demystifying-precision-and-recall-in-machine-learning-6f756a4c54ac>.

Any Questions?



Kaplan Business School
T2 2024

Thank You!


Source: Unsplash