# CHAPTER III

# RESEARCH METHODOLOGY

## 3.1 Research Object

The research objects in this study are the image data of three common skin cancer types, that is BCC (Basal Cell Carcinoma), SCC (Squamous Cell Carcinoma), and Melanoma. According to (Thissen et al., 2017), each of these skin cancer types has it is characteristic, even though the root cause is usually due to long and intense exposure to the sun's UV radiation. The characteristic differences can be viewed in Table 3.1.

**Table 3.1 Skin Cancer Object Information**

| Skin Cancer Type | Growth Location | Symptoms |
|---|---|---|
| BCC (Basal Cell Carcinoma) | Skin area that are usually exposed to sun such as **face, neck, ears, scalp, shoulders, and back.** | 1. Raised, smooth, and pearly bump<br>2. Visible small blood vessels<br>3. Bleeding (ulceration) inside the central part of the tumor |
| SCC (Squamous Cell Carcinoma) | Skin area that are usually exposed to sun and showing signs of sun damage such as **wrinkles or age spots** | 1. Red and scaly thickened bump<br>2. Chance of bleeding (ulceration) inside the tumor |
| Melanoma | Any area of the body | 1. Tumor with a mixture of black, tan, white, blue, red, and brown color<br>2. Asymmetry shape between two sides of the tumor<br>3. Irregular tumor's borders line |

In this research, the data that will be used is skin cancer images data collected from ISIC (International Skin Imaging Collaboration). The images data consist of 3 skin cancer type with a total data of 8669 images. Each image has a standardized

format, the skin cancer is located in the middle with the surrounding skin covering the rest of the image. This standardized format means that each image has no background distraction. The images are also taken with good lighting which ensures clear visibility. A sample image of basal cell carcinoma (BCC), squamous cell carcinoma (SCC), and melanoma can be viewed in Image 3.1.

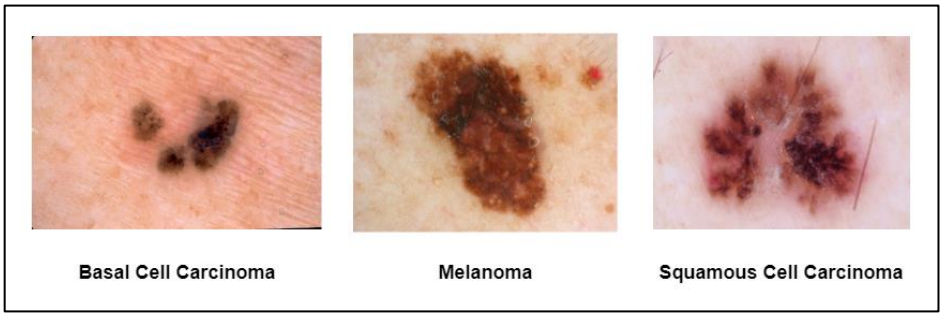


**Image 3.1 Common Skin Cancer Image Sample**

## 3.2 Research Method

### 3.2.1 Classification Method

Due to the nature of this research, the ideal classification method would have a good track record of achieving high image-based classification accuracy and has great flexibility so that it can be adjusted accordingly to suit the research's needs. Taking this into account, Table 3.2 show the comparison of several classification methods that can achieve this goal [14].

**Table 3.2 Classification Method Comparison Table**

| Category | CNN (Convolutional Neural Network) | SVM (Support Vector Machine) |
|---|---|---|
| Common use-case | CNN is a non-linear classifier that works well with visual or images-based recognition | SVM is a linear classifier that works well with common classification problem |

| | | |
|---|---|---|
| Robustness and Configurability | CNN can adjust (increase or decrease) their own model complexity by adding more layer and tuning the layer parameters | SVM does not have the ability to adjust their own model complexity |
| Documentation | CNN has a comprehensive and detailed documentation for image-based related work | SVM has a more well-rounded and less detailed documentation to support all types of classification |
| Computing Resource | CNN require higher number of resources to train compared to other method due to complexity reason. However, CNN is well-tailored to take advantage of powerful hardware such as GPU (Graphical Processing Unit) | SVM require lower number of resources to train compared to CNN. SVM also does not support the usage of GPU (Graphical Processing Unit) |

Based on the classification method comparison in Table 3.2, CNN (Convolutional Neural Network) algorithm will be used to classify the types of skin cancer based on the cancer characteristic in the image. The CNN's flexibility and robustness will be able to support the large and feature-rich dataset of the skin cancer images. Furthermore, the CNN's algorithm will also be able to utilize the available GPU (Graphical Processing Unit) that this research possesses.

### 3.2.2 Proposed Method

CNN's algorithm will be used to power the skin cancer type classification model. An accurate CNN classification model requires proper planning and preparation before the actual completed model is ready to be used to classify skin cancer type in a web-based application. In similar research, the CRISP-DM (Cross-Industry Standard Process for Data Mining)

framework was used as a base method on preparing a CNN model that is capable of classifying ulos images. Using this method, the research can achieve a favorable and robust CNN model [31]. Taking this into account, preparation steps in this research will be adjusted according to the steps in CRISP-DM which can be viewed in Image 3.2.

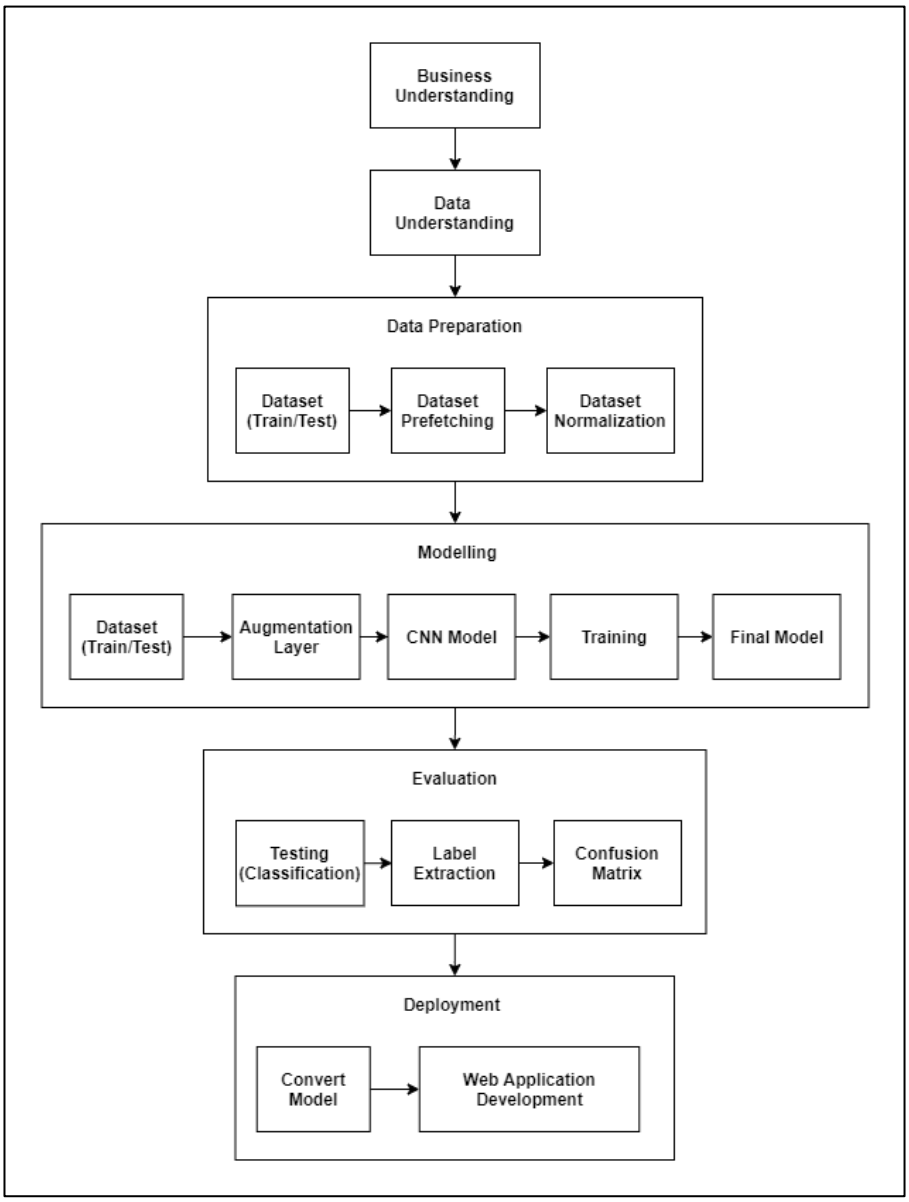


**Image 3.2 CRISP-DM Implementation in Classification Model Creation**

Based on Image 3.2, the stage that is adjusted according to CRISP-DM steps are :

1. Business Understanding

   In this step, the process is identifying the root problem and the creation purpose of the image-based skin cancer type classification in a web-based application using CNN Algorithm. This image-based approach has also been implemented on other things, such as lumbar spinal stenosis detection, kidney detection, and others.

2. Data Understanding

   In this step, the process is about breaking down the data acquisition process, starting from the source and the provider's validity. This step will also describe the dataset's general information and data quality in detail.

3. Data Preparation

   In this step, the dataset will be prepared for the training process to ensure that the CNN's model will be able to fully utilize the raw data. There are various preparations in this step, such as :

   a. Loading the dataset to a variable. Since the dataset is collected manually from ISIC's website, the dataset is manually separated into train and test folder with a ratio of 8:2.

   b. Prefetching the dataset before the training process begins will fasten the training process. This process is about delegating the

**feeder** role to the CPU (Central Processing Unit) while the GPU (Graphical Processing Unit) can keep focusing on the model training process or the consumer role.

c. The dataset normalization process is about converting the 0-255 value that is commonly used to represent image data into a 0-1 decimal value. This process will greatly reduce the training speed since there is a smaller range of number to be processed.

4. Modelling

In this step, the prepared data is finally ready to be fed to the skin cancer type classification model. The CNN's model consists of 2 components, that are :

a. Augmentation Layer

This layer will intercept the input data and perform an augmentation on the image before being forwarded to the main CNN model. This step will ensure more variations of the input data are being fed to the main model. With more variations, the model will be able to better generalize the data to prevent overfitting. The augmentation transformation was achieved by using the preprocessing layers that were provided by the library Keras. Keras is a very popular library for neural networks or other machine learning model functions and was used by many research papers such as [8][24][31][32]. Keras

provided several preprocessing layers for augmentation purposes, but this research will only use 3 layers, that is, random flip, random rotation, and random zoom due to memory limitations.

b. CNN Model

In this main model, the model will extract and learn the skin cancer feature from the images.

5. Evaluation

In this step, a performance evaluation of the trained model will be conducted by doing a separate classification process using the test data. By comparing the predicted result of the model and the true label of the image data, a confusion matrix can be created to further analyze the model's ability to perform classification based on each label.

6. Deployment

In this step, a web-based application that can accept image input and output the probability result of the classification will be created. The final and evaluated model will be used as the system backbone.

## 3.3 Research Variable

Research variables are separated into two types, dependent variable, and independent variable. In this research, there are 1 dependent variable and 3

independent variables which can be viewed in Image 3.2. The dependent variable in this research is skin cancer type (Y) while the independent variable is color (X1), shape (X2), and pattern (X3).

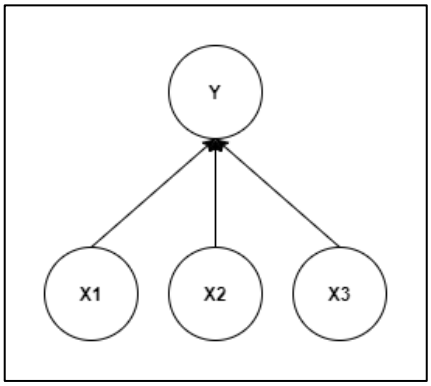


**Image 3.3 Research Variable**

## 3.4 Tools

### 3.4.1 Programming Language

This research will use Python as the programming language to create the skin cancer type classification model. In the context of image classification, Python has unparalleled library support and high usability compared to another language which can be viewed in Table 3.3.

**Table 3.3 Programming Language Comparison Table**

| Category | Python | Java | R |
|---|---|---|---|
| Usability | Python is the most popular programming language for any data-science-related workload. Python is extremely user friendly due to the simple syntax and code readability | Java is the go-to language for application building but can also be used for model building even though the usage is extremely rare. Java has more complex syntax and not as easy to understand as | R is a popular data science language that leans more towards statistics and structured dataset |

| | | | |
|---|---|---|---|
| | | other data science language | |
| Libraries | Python has many libraries dedicated to data-science related workload like Keras, Tensorflow, etc | Little to no-existent libraries for data science | R has many libraries for data science-related workload |
| Performance | Due to the interpreter language nature, Python is slower than other compiled language | Java is efficient and fast because of the compiled language nature | R is an interpreter language, slower than other language with compiled nature |

In this research, there are several Python's main module that are being used in the classification model creation which can be viewed in Table 3.4.

**Table 3.4 Python's Main Module in This Research**

| Module | Purpose |
|---|---|
| TensorFlow | A library that supports image preprocessing stage, such as, loading dataset via a batch system or performing data augmentation, and providing model creation tools |
| NumPy | A library that supports data manipulation |
| MatPlotLib | A library that supports data visualization in various form |
| Tensorflowjs_converter | A library that converts a .h5 model file into a JavaScript compatible format |

### 3.4.2 Hardware

In this research, all computing operation is performed in a local environment with a specification of

a. Intel i7-8750H (CPU)

b. 16GB RAM

c. NVIDIA GTX 1050Ti (GPU)

### 3.4.3 Software

In this research, all software that is used to support the classification model creation are as follow :

a. JupyterLab that acts as a code editor for the Python language in the classification model creation.
b. Visual Studio Code that acts as a code editor for the HTML, CSS, and JavaScript language in web application development.
c. Anaconda Navigator manages the Python's module and version in one place.