CHAPTER II

THEORITICAL FRAMEWORK

2.1. Skin Cancer

Skin cancer is caused by out-of-control abnormal cell growth in the outermost skin. This can happen due to unrepaired DNA damage that triggers mutations. These mutations trigger the abnormal cells to multiply rapidly, which results in a form of malignant tumors. There is a lot of variety regarding skin cancer, each with it is unique causes and danger rate. However, the general rule of thumb is every skin cancer is curable if detected early because the cancer cells have yet to grow bigger or worse, spreading to other parts of the body. Every skin cancer can be deadly/fatal in the long run if left without treatment. There are 3 types of the most common skin cancer in the world, that are, BCC (Basal Cell Carcinoma), SCC (Squamous Cell Carcinoma), and Melanoma [18].

2.1.1. BCC (Basal Cell Carcinoma)

BCC (Basal Cell Carcinoma) is a result of over-exposure to ultraviolet (UV) radiation. BCC (Basal Cell Carcinoma) occurs when the outermost layer of human skin or epidermis suffers DNA damage from over-exposure to UV radiation. The damaged basal cell will then start to grow uncontrolled. Usually, BCC (Basal Cell Carcinoma) will not spread to another part of the human’s body, however, if this tumor is not treated as soon as possible, the untreated BCC (Basal Cell Carcinoma) can spread deep into the skin, tissue, and even bone. Potentially disfiguring the area of the BCC (Basal Cell Carcinoma) permanently. In a rare instance of BCC (Basal Cell Carcinoma) spread to another part of the body, however, this aggressive cell can be very life-threatening. BCC (Basal Cell Carcinoma) itself can appear differently from a human to another human, however, the general rule of thumb is that BCC (Basal Cell Carcinoma) usually looks like pink or red shiny bumps with a slightly elevated growth. Sometimes, BCC (Basal Cell Carcinoma) may even bleed or itch [19].

2.1.2. SCC (Squamous Cell Carcinoma)

SCC (Squamous Cell Carcinoma) is also a result of over-exposure to ultraviolet (UV) light radiation. The squamous cell, a cell located in the outermost layer of a human's skin is very active as it continuously sheds new form. When damaged from over-exposure to UV radiation, this cell will also grow uncontrollably like BCC (Basal Cell Carcinoma) cell. The majority of SCC (Squamous Cell Carcinoma) cell can be easily treated by a dermatologist, however, if this cell is allowed to grow over some time, the SCC (Squamous Cell Carcinoma) cell can penetrate deep inside the skin and spread to another part of human's body. The SCC (Squamous Cell Carcinoma) cell also looks unique on every human being but usually, this cell can appear as a thick or rough skin that may have a crust over it, or even bleed or itch [20].

2.1.3. Melanoma

Melanoma is the least common type of skin cancer compared to BCC (Basal Cell Carcinoma) or SCC (Squamous Cell Carcinoma). However, melanoma is the most dangerous and aggressive type of skin cancer out of the three, as it can spread deep inside the body and reach the human body organ. Melanoma is a term for an uncontrollable growth of Melanocytes cells. Melanocytes cell is located in the upper part of human’s skin which function as pigment producer, the reason human’s skin has a different color. Melanoma is also triggered by over-exposure to UV light radiation, the UV light triggers the Melanocytes cell to produce more pigment which causes a mutation in the Melanocytes cell itself. The mutated Melanocytes cell will grow uncontrollably if left untreated. Melanoma cell is curable, provided that the cell is detected and treated at an early stage. However, Melanoma usually appears in different sizes, color, and shapes which makes it very challenging to be detected [21].

2.2. CRISP-DM (Cross-Industry Standard Process for Data Mining)

CRISP-DM is an international standard used for a data-mining related task, this framework consists of 6 phases, such as business understanding, data understanding, data preparation, modeling, evaluation, and deployment [22]. The explanation for each of these steps are as follow :

1. Business Understanding

Business Understanding is the first step with the purpose of understanding the business goals and objectives, understanding the business context or situation, and from this point, determining how to connect this business information directly to the data mining goals. [22]

2. Data Understanding

   Data Understanding is the second step with the purpose of understanding the data acquisition phase. The purpose of this step is understanding the data, determining the data quality, and finding interesting anomaly from the pattern of the data. [22]

3. Data Preparation

   Data Preparation is the third step with the purpose of determining the data that will be used for the next step and also removing any defects, outliers, from the data. [22]

4. Modeling

   Modeling is the fourth step with the purpose of determining the correct data mining techniques to be used by looking at the research variable and the initial goals of the data mining goals. [22]

5. Evaluation

   Evaluation is the fifth step with the purpose of checking the result by comparing the actual result against the defined business goals or objectives. [23]

6. Deployment

Deployment is the final step with the purpose of finalizing the result, either in the final report or another software related program. Other than deploying, the tasks of monitoring and maintaining are also equally important. [23]

The complete stages of the CRISP-DM standard can be viewed in Image 2.1.

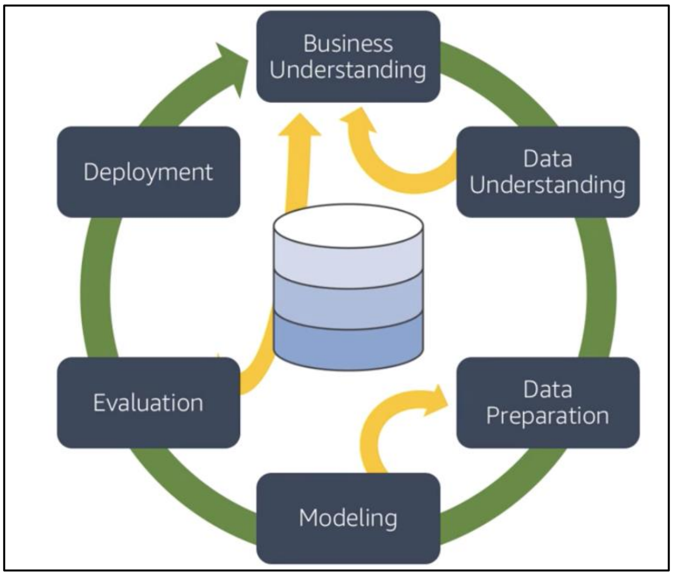


Image 2.1 CRISP-DM Cycles

2.3. Convolutional Neural Network (CNN)

2.3.1. CNN Algorithm

Convolutional Neural Network (CNN) is a widely used deep learning architecture for image or visual-related tasks such as object recognition, image classification, or others. CNN (Convolutional Neural Network) is based on MLP or Multi-Layer Perceptron algorithm. The key difference between these two algorithms is that the CNN nodes are only connected to some nodes in the preceding layers while MLP is fully connected with all the nodes in the preceding layers. This small difference allowed CNN to perform better at an image-related workload. Another ability of the CNN model is that the algorithm is designed to be able to process more than one-dimensional data. [24]

2.3.2. CNN Layer

There is various type of layer that exist and commonly used in the CNN (Convolutional Neural Network) algorithm, such as :

1. Convolutional Layer

   Convolutional Layer is a layer that main purpose is to extract as much information as possible from the input data. [25]

2. Pooling Layer

   Pooling Layer is usually the next layer after the Convolutional Layer since the function of the Pooling Layer is to lower the overall complexity of the extracted feature by keeping only the highest value (Max Pooling) or only keeping the average value (Average Pooling). [25]

3. Fully Connected Layer or Dense Layer

   A Fully Connected or Dense Layer is the final most layer, where all of the information from the previous neuron or layer end at this layer. Using this information, the Fully Connected or Dense Layer will then determine the label of the input data by outputting probabilities with a sum of 1. [25]

4. Dropout Layer

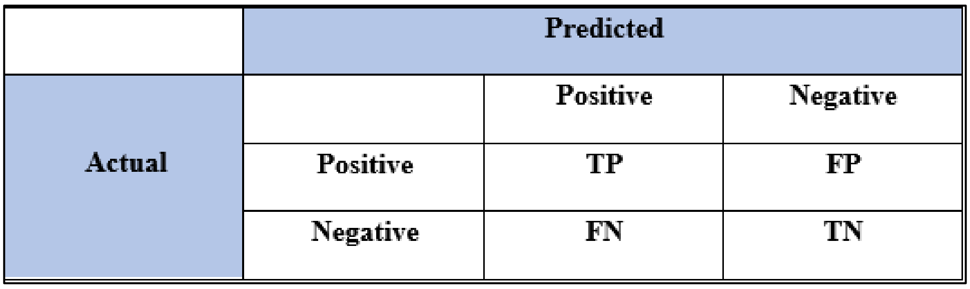
Dropout Layer’s main purpose is to improve the model’s ability to regularization and to prevent overfitting. The Dropout Layer works by randomly setting a neuron value in a layer into 0, which fasten the overall training process of the model due to lower data being processed. [25]

2.4. Data Augmentation

Data Augmentation is a widely popular technique used to increase the overall data variety that is being fed to the model training. Data Augmentation works by artificially generate another variant of image data based on the original image by performing a transformation on it, such as scaling, cropping, noise injection, color space transformation, rotations, horizontal flipping, vertical flipping, and others. With this approach, data augmentation can boost an already big dataset into having more data variety, and in the case of a small dataset or dataset that have an imbalance label where one label data size is extremely small compared to the other label, data augmentation can help by generating more image so that the model can learn better in the training process. [26]

2.5. Confusion Matrix

A confusion matrix is one of the most popular methods for model evaluation that focuses on binary classification or multiclass classification. The confusion matrix works by comparing the predicted result with the actual value of the data. Based on this comparison method, the confusion matrix generates 4 additional terms, that is, True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). True Positive (TP) represents the number of predicted results that are correctly classified as positive. True Negative (TN) represents the number of predicted results that are correctly classified as negative. False Positive (FP) represents the number of negative data incorrectly classified as positive. False Negative (FN) represents the number of positive data incorrectly classified as negative. The confusion matrix table can be viewed in Image 2.2. [27]



	Predicted		
Actual		Positive	Negative
	Positive	TP	FP
	Negative	FN	TN

Image 2.2 Confusion Matrix Table

Based on Image 2.2, the accuracy of a classification model can be measured using Formula 2.1. [27]

$$\boxed{\textit{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}}$$

Formula 2.1 Accuracy Measurement using Confusion Matrix

2.6. Tools

2.6.1. Python

Python programming language is an open-source interpreter language that can utilize object-oriented programming, procedural style programming, and others. In recent years, Python is very popular and widely known as the go-to language for data science [28]. This is due to various factors, such as :

1. Versatility

The programming language of Python can be run on any modern days operating systems such as Windows, macOS, and Unix. Python is also a very lightweight programming language to run. These two advantages enabled Python to be run virtually everywhere on any device. [28]

2. Extensive Library and Documentation

Python has one of the most well-detailed documentation and library support in the history of programming language to cover any type of usage, ranging from simple tasks such as data manipulation up to deep learning model creation. Due to the open-source nature of the library and documentation, the documentation and library are constantly under regular update by the community. [28]

3. User Friendly

Python has a very small learning curve due to the fact the language is very similar to human language. Combined with the extensive library and documentation, the Python language can be easily used even by people that have no prior knowledge of programming. [28]

2.6.2. TensorFlow

TensorFlow is one of Python’s most popular libraries due to the robustness of the library in handling any data science-related stuff. TensorFlow is a machine learning framework, originally developed by Google, to support the end-to-end machine learning process starting from the dataset preparation phase up to the actual model deployment into the application. Furthermore, since TensorFlow is initially geared towards big scale machine learning process, this library is also capable of doing parallel processing using only CPU/GPU setup or even multi-GPU setup. [29]

2.7. Previous Research

Table 2.1 Previous Research Analysis

No	Journal	Authors	Research Result	Contribution in This Research
1	IEEE Access. Boundary Delineation of MRI Images for Lumbar Spinal Stenosis Detection Through Semantic Segmentation Using Deep Neural Networks. 7, 43487-43501. [15]	Al-Kafri, Ala S. Sudirman, Sud Hussain, Abir Al-Jumeily, Dhiya Natalia, Friska Meidia, Hira Afriliana, Nunik Al-Rashdan, Wasfi Bashtawi, Mohammad Al-Jumaily, Mohammed	The result of this study is a robust semantic segmentation with CNN model that is able to detect stenosis in lumbar spinal using MRI images.	Contributing as reference that deep learning CNN algorithm are much better for modern day image classification issue like skin cancer type detection compared to older algorithm like SVM to achieve better efficiency and performance accuracy.

2	Journal of King Saud University - Computer and Information Sciences. Intelligent handwritten recognition using hybrid CNN architectures based-SVM classifier with dropout. 1-7. [14]	Ali, Amani Ali Ahmed Mallaiah, Suresha	The result of this study is a hybrid SVM-CNN model that can recognize handwritten Arabic data with a favorable result when compared to the state-of-the-art Arabic text recognition model.	Contributing as reference to utilize dropout layer to prevent overfitting and to use CNN as model algorithm due to excellent auto feature extraction
3	Pattern Recognition Letters. Data augmentation method for improving the accuracy of human pose estimation with cropped images. 136, 244-250. [30]	Park, Soonchan Lee, Sang baek Park, Jinah	The result of this study is that the use of data augmentation can increase average accuracies without the any modification on the neural network architecture.	Contributing as reference to utilize data augmentation technique to improve overall model accuracy
4	European Journal of Cancer. Deep neural networks are superior to dermatologists in melanoma image classification. 119, 11-17. [16]	Brinker, Titus J. Hekler, A. Enk, Alexander H. Berking, Carola Haferkamp, Sebastian Hauschild, Axel Weichenthal, Michael Klode, Joachim Schadendorf, Dirk Holland-Letz, Tim	The result of this study is a CNN based model that is able to do a melanoma and nevi prediction better than certified dermatologist.	Contributing as reference to the downside of using CNN model without the support of pre-trained model poses little to no difference.

		von Kalle, Christof Fröhling, Stefan Schilling, Bastian Utikal, Jochen S.		

The comparison between previous research and this research are :

1. The first research in Table 2.1 uses semantic segmentation with CNN to develop the model while this research uses conventional CNN.
2. The second research in Table 2.2 uses a hybrid SVM-CNN model while this research uses a standalone CNN model.
3. The third research in Table 2.3 uses a Neural Network algorithm while this research uses a CNN algorithm
4. The fourth research in Table 2.4 uses Melanoma and Nevi as their research object while this research use BCC (Basal Cell Carcinoma), SCC (Squamous Cell Carcinoma), and Melanoma.
5. The fourth research in Table 2.4 share the same research topic with this research, that is, regarding dangerous skin cancer conditions. However, while the fourth research final output is a model. This research took it further by creating the final output of a web-based application that can detect skin cancer type (BCC, SCC, Melanoma) using image data input.