

**How to improve the
Performance of Skin
Cancer detection using a
Machine Learning
approach such as
Convolutional Neural
Network.**

Assignment 3: Full Research Proposal

Student: Vinay Huchanahalli Nagaraju.
(N10180893)

Supervisor: Dr. Patrick Delaney

Contents

Problem Statement.....	2
Research Question:	2
Research Methodology:.....	3
Expected Tangible Outputs	7
New Knowledge:	8
References:	10
Reflective Statement:	10

Abbreviations:

CNN – Convolutional Neural Network
UV – Ultraviolet
GPU – Graphical Processing Unit
CPU – Central Processing Unit
AUC - Area under the Curve
ROC - Receiver Operating Characteristics
TP - True Positive
TN - True Negative
FP - False Positive
FN - False Negative
EDA – Exploratory Data Analysis

How to improve the Performance of Skin Cancer detection using a Machine Learning approach such as Convolutional Neural Network.

Problem Statement

Skin cancer is one of the most diagnosed types of cancer in Australia. Skin cancer is primarily caused by excessive exposure to ultraviolet (UV) radiation from the sun. This will cause DNA damage to skin cells. If the damaged DNA doesn't recover by itself then that will cause abnormal growth of cells. Which eventually became skin cancer. There are many types of skin cancer, in which Melanoma is dangerous one. Melanoma is most dangerous skin cancer type which kills more than 1,700 lives in 2016 (Shih, Carter, Heward, & Sinclair, 2017)ⁱ. Squamous cells carcinoma and basal cell carcinoma are also threatening but they are not that much dangerous when compare to Melanoma. If Melanoma is identify in early stage then it is easy to save thousands of lives in Australia. Convolution Neural Network (CNN) will help to identify different types of skin cancer and Melanoma is one among them. CNN will scan the damage part of skin image. Initially, CNN will try to detect edge of skin pattern by using a technique called kernel or filter (Esteva & Kuprel, 2017)ⁱⁱ. CNN is a machine learning model which follows "Supervised Learning" approach. Supervised learning means machine will learn from previous cases. There will be Target variable which is present in "Supervised Learning". If the patient is having skin cancer then Target variable is 1 else Target variable will be 0. We can import dataset from "Kaggle" website but, it is not accurate and there are many errors present in the dataset such as Typing errors, outliers, etc. (Pham, Luong, 2018)ⁱⁱⁱ. Kernel will sometimes compress the image of skin cancer. Hence, there is a chance of information loss in skin patterns. These errors will impact the performance of the CNN model. Hence, we need to develop our own framework to collect the data in Queensland hospitals. By considering the above problems, it is necessary to improve data collection process and kernel or filter technique to improve the performance of skin cancer detection by using Convolutional Neural Network model.

Research Question:

The main objective associated with this Literature is to improve the kernel or filter. Because, kernel is mainly used in the CNN model to detect skin cancer. There are many factors which affects the performance of kernel they are missing values which are present in Data, Blur image, computational speed of computer, and many other factors. We need to find the solution to improve the data quality and that will enhance the performance of kernel. I found 3 research questions to enhance the performance of Convolutional Neural Network, but, finding solution to below problem will help to enhance the cancer detection and it is more appropriate to my main research problem.

"What enhancements will improve image classification performance in Convolutional Neural Network model to detect skin cancer in Queensland?"

This research problem cannot be resolved using the existing system because, by default we are not using kernel and there are many problems which are associated with the quality of

data. If the quality of the data is not good then kernel will not perform better. We are mainly using kernel to detect skin cancer image or to analyse skin cancer pattern. If kernel fails to analyse skin cancer pattern then CNN model fails to recognise the cancer accurately (Brinker & Hekler, 2018)^{iv}. We need to identify the importance of kernel in detection of skin cancer. We need to apply various technique to improve the performance of kernel.

If the performance of kernel improves then we can apply this CNN model with Kernel on real world cases. Which will reduce the user time to diagnose skin cancer. It will also reduce the waiting time. Because CNN model will detect skin cancer in less time. Hence, finding solution to this question will impact a lot in a real world. Hypothesis testing can be used to determine the importance of kernel.

Research Methodology:

This research is artefact oriented approach that will develop a Machine learning model to detect skin cancer in early stage by using Convolutional Neural Network. It is a prerequisite to read all research article or literatures which are related to CNN and skin cancer. Hence, it is considered as stage 0. Research Methodology mainly consists of 5 steps, they are:

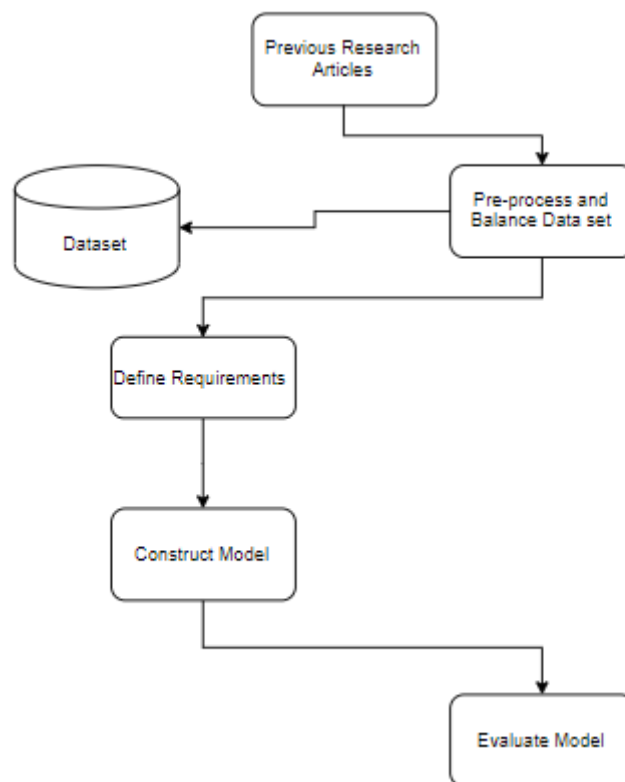


Figure 1: Steps involved in Project Methodology

Stage1: Pre-process and Balance Dataset

Data which is present in “Kaggle” or in real world is filled with many errors. We need to reduce these errors in order to improve the performance of CNN. In Data Science, the process of removing noise, typo errors, or outliers is called as Imputation. Another method is to balance the dataset. Machine learning models such as CNN does not work well with imbalanced dataset. Imbalanced data is having bias or inclined towards one classification (Fotouh, Asadi, & Kattan, 2019)^v. In this method we are using Binary classifier. If the skin

cancer is present then we can classify that image as 1 else we can classify the image as 0. Below is the image which represents the classification of skin cancer.




Image	Classification
	1
	0
	0

Table 1: Dataset

Stage 2: Define Requirement

There are 3 types of requirements which is associated with this project, they are:

1. Human Resource:
 - 1.1. This Research needs at least 2 masters student and 1 Doctorate student.
 - 1.2. We need master students to collect skin cancer image in Queensland.
 - 1.3. We need Doctorate student to improve the performance of CNN by altering kernel or filter.
2. Computational Requirement:
 - 2.1. Size of RAM must be greater than or equals to 16 GB. Because, CNN is a complex model which needs high RAM for computation.
 - 2.2. GPU is needed to run Convolutional Neural Network. Because, normal CPU is having very less number of cores with high memory on each cores. But, CNN need more cores with less memory on each cores.
 - 2.3. There are different types of GPU present in the market. NVIDIA GTX 1080 is better for the computation of CNN.
3. Software and Programming:

Integrated Developing Environment (IDE): Spyder, Pycharm or Jupyter Notebook.

Programming Language: Python Programming Language (Version 3).

Libraries:

Task	Libraries
Data Pre-process or Imputation	Sklearn, pandas, numpy.

Computation of CNN	Keras, TensorFlow, sklearn.
Data Visualization	Matplotlib, seaborn.

Table 2: Python Libraries

Stage 3: Constructing Machine Learning Model (CNN)

Convolutional Neural Network (CNN) is the Machine Learning model that we are using to detect skin cancer. CNN consists of 2 steps

Step 1: Feature Extraction:

In this step, Feature of skin is extracted from the image. Kernels or filters are used to detect the patterns of skin cancer. Kernel is present between Convolution layer and pooling layer. While extracting the features, kernel will shrink the image. Hence, these leads to the loss of information of skin cancer. Pooling layer will replicate or duplicate the image for future pattern extraction.

Step 2: Classification:

We are using fast forward neural network to classify the image. In Hidden layer, each node is consider as perceptron. Hence, this is also called as Multiple Layer Perceptron (MLP). Each node is connected to all the other node in next layer. Hence, Hidden layer is fully connected. Each node is having some activation function. Which will activate the node based on the skin cancer image. Output layer is having 2 nodes. One node represents Skin cancer and another node represents not a skin cancer. For example:

Node 1: 0.89977 (Skin cancer present)

Node 0: 0.21133 (Skin cancer absent)

Node 1 is having highest probability. Hence, CNN will classify that the patient is having skin cancer. Below diagram illustrates the classification of Convolutional Neural Network.

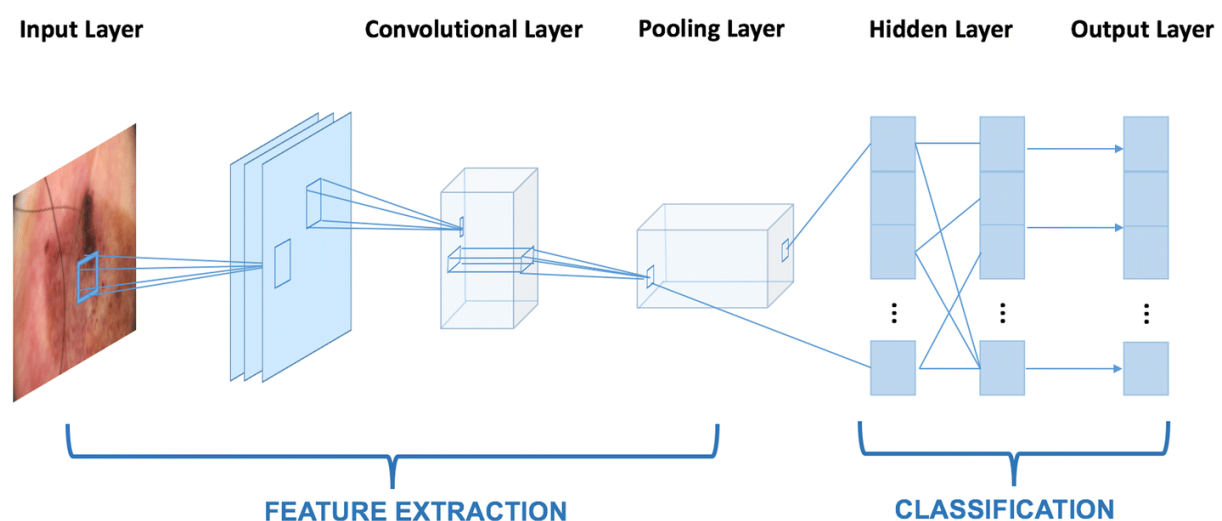


Figure2: Source:

<https://bmcmmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-018-0631-9>

Step 4: Evaluation of Model

Kernel plays vital role in the classification of skin cancer. We are using 2 types of evaluation model to estimate the importance of kernel in Convolutional Neural Network.

Method1: Hypothesis Testing:

In Hypothesis Testing, we are considering Null hypothesis and Alternate Hypothesis. If the kernel or filter is having significant level in determining or enhancing skin cancer detection then alternative hypothesis must be true. Null hypothesis is represented as H_0 . Initially, we need to feed the Data to CNN and based on the accuracy of CNN with or without Kernel, we can either accept the null hypothesis or reject the null hypothesis.

Null Hypothesis: Kernel is not having any impact on detecting skin cancer using CNN.

Alternative Hypothesis: Kernel is having some impact on detecting skin cancer Using.

Decision is:	The Null Hypothesis is	
	True	False
Accept H_0	$(1-\alpha)$ Confidence Level	β
Reject H_0	α	$(1-\beta)$ Power of the test

Image1: source: "<https://businessjargons.com/hypothesistesting.html>"

Figure 3: Hypothesis Testing

Method 2: AUC – ROC Curve:

This is another method to identify the importance of Kernel. AUC (Area under the Curve) ROC (Receiver Operating Characteristics) curve is used in binary classification. It is one of the important evaluation for checking any classification model's performance. We are using binary classification to predict the importance of kernel.

Binary classification – 0: if the patient is not having skin cancer
1: if the patient is having skin cancer

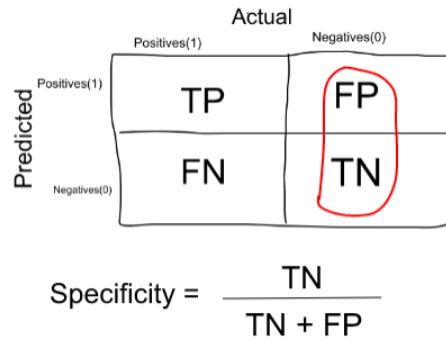


Figure 4: Confusion Matrix for AUC ROC curve

True Positive (TP): If the actual and Predication from CNN model is same.

True Negative (TN): If the actual and Predication from CNN model is not same.

False Positive (FP): If the actual and Predication from CNN model is not same.

False Negative (FN): If the actual and Predication from CNN model is same.

When we decrease the threshold, we get more positive values thus increasing the sensitivity. Meanwhile, this will decrease the specificity.

Data analysis

In CNN, Exploratory Data Analysis (EDA) is an approach to analyse the given data sets to summarize their main characteristics. We can take each features and we can analyse the pattern of skin cancer.

Expected Tangible Outputs

We need to create a framework to collect data from Queensland hospitals. We can ask hospitals receptionist to capture the image of skin cancer and store it in their local computers. Every week we insist our master students to collect those data from all hospitals in Queensland. We need at least 50,000 images to train our CNN model. We need to add these additional images into a dataset in which we derived from 'Kaggle'. If dataset is more then we can train our model well. After collecting those number of images, we need to evaluate the importance of Kernel. We need to divide the dataset into two parts. First part is considered as Data A and second part is named as Data B. For Example, if the total number of image is 50,000 then Data A will have 25,000 data and Data B will have 25,000 data. We are choosing the dataset randomly.

Data	Number of Data
Data A	25,000
Data B	25,000

Table 3: sample Dataset

Data A is using Kernel or Filter to detect skin cancer using CNN and Data B is not using any kernel to detect skin cancer using CNN. Then Data A and Data B is trained and tested separately using Convolutional Neural Network. Accuracy obtained from CNN model is:

Data	Number of Data	Accuracy	AUC ROC Curve
Data A	25,000	88.4%	0.86
Data B	25,000	76.2%	0.641

Table 4: Performance of Model

Hence, we can predict that kernel plays major role in the detection of skin cancer using CNN machine learning model.

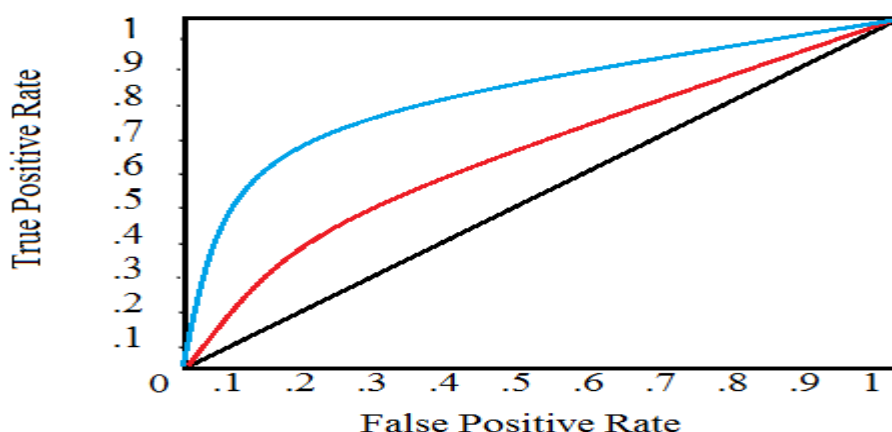


Figure 5: AUC ROC curve

Above image drawn using AUC ROC curve. Blue mark indicates the performance of CNN using Kernel or Filter. Red curve indicates the performance of CNN without using AUC ROC curve.

By the above method, it is confirmed that filter or kernel plays a major role in detection of skin cancer. Hence, we need to create a framework to make the usage of kernel or filter in Convolutional Neural Network. We need to consider set of customers who is diagnosed with cancer. Hence, we can reject null hypothesis and declare that Kernel is enhance the performance of CNN model in detecting skin cancer.

We can use this research proposal for Publications and printed material. This will help to give the beginners a brief idea about CNN and its implication on skin cancer. These outputs will have benefits for medical practitioners such as doctors or nurse to identify the cancer in early stage. Early detection of skin cancer will help to save many lives.

New Knowledge:

After conducting the experiments, we understand the importance of kernel in predicting skin cancer by using Convolutional Neural Network. We are getting 88.4% accuracy after implementing kernel and 76% accuracy without using kernel. There are many steps to further improve the accuracy of model by improving the kernel in CNN model they are:

1. We can make kernel as a default parameter in CNN model to increase skin cancer detection.
2. We can implement this technique to other cancer detection models and check their accuracy.
3. We can also identify various other method to increase the performance of kernel or filter in CNN model.

By using AUC ROC curve, we can determine the importance of Kernel. This new knowledge leads to further development of other techniques which work better than kernel. This new knowledge will also help to enhance the performance in detection of other types of skin cancer such as recurrent basal cell carcinoma, Squamous cell carcinoma, Merkel cell carcinoma, etc. This type of skin cancer are not fatal but they may cause serious damage to the skin. This new knowledge can be applied to detect Siamese manta rays fish.

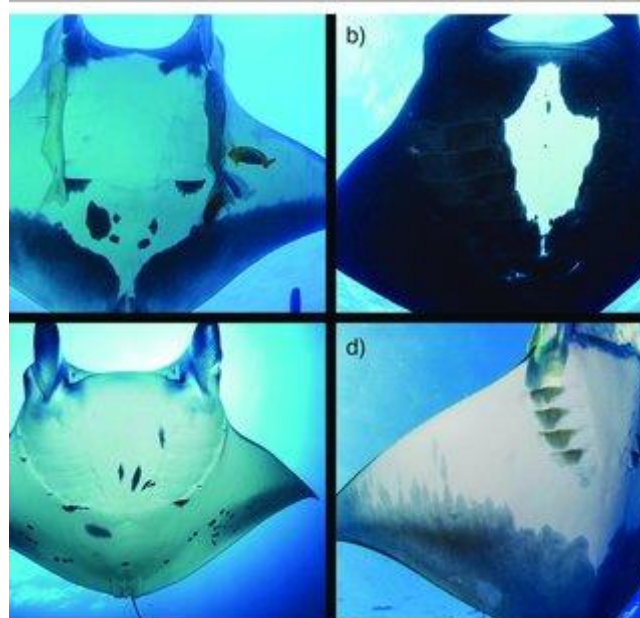


Figure 6: Source: "<https://arxiv.org/pdf/1902.10847.pdf>"

Overall, Kernel will enhance the performance of skin cancer detection. It is not always optimal because, it shrinks the image and cause information loss. This impact does not affect much on skin cancer image but, this little impact will not make us to reach more accuracy in skin cancer detection. Hence, we need to research more on kernel and find possible ways to improve it or we need to find alternative approach or technique which works better than kernel.

Researcher can use this knowledge to save the manta ray fish in Blue-reef, Queensland. Due to tourism, number of manta ray fish is diminishing. Researcher can use this new knowledge to identify those fish and they can recommend new system or methodology to save them.

We can consider various other fields such as Forensic science to identify criminals. We can use this new knowledge along with Siamese Network in Convolutional Neural Network to identify criminals.

References:

- ⁱ Shih, S., Carter, R., Heward, S., & Sinclair, C. (2017). Economic evaluation of future skin cancer prevention in Australia. *Preventive Medicine*, 99, 7–12.
- ⁱⁱ Esteva, A., Kuprel, B., Novoa, R., Ko, J., Swetter, S., Blau, H., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks.
- ⁱⁱⁱ Pham, T., Luong, C., Visani, M., & Hoang, V. (2018). Deep CNN and Data Augmentation for Skin Lesion Classification (Vol. 10752, pp. 573–582).
- ^{iv} Brinker, T., Hekler, A., Utikal, J., Grabe, N., Schadendorf, D., Klode, J., ... Von Kalle, C. (2018). Skin Cancer Classification Using Convolutional Neural Networks: Systematic Review.
- ^v Fotouhi, S., Asadi, S., & Kattan, M. (2019). A comprehensive data level analysis for cancer diagnosis on imbalanced data.

Reflective Statement:

Student: Vinay Huchanahalli Nagaraju

Supervisor: Dr. Patrick Delaney

Week	Supervisor	Student
11	Patrick explained about reframing the research questions. He taught about Qualitative, Qualitative and Artefact oriented research. He explained how to do the Methodology and things that he is expecting in the Project Methodology.	I understood the method to reframe my research question. I understood various types Research Methodology such as Qualitative, Qualitative and Artefact oriented. Since, my research methodology is artefact oriented, I gave more importance to artefact oriented approach.
12	He taught the various ways to analyse the data. He explained how to conduct the tangible output and how to explore new knowledge by using Project Methodology. He also took one example and taught how to explore new knowledge and tangible output using artefact oriented approach.	I initially found many challenges such as finding artefact oriented approach. But, Patrick made it clear in week 12. So, understood the procedure to explore new knowledge and analysis of tangible output.
13	He went through all section briefly and he gave more priority to research methodology. He again explained about Qualitative, Qualitative and Artefact oriented research approach. He also went through the assignment 3 and explained the things he is expecting in Assignment 3.	I understood the marking criteria for assignment 3. I understood the artefact oriented research approach in depth. I showed my assignment 3 draft to Patrick for his feedback. He analysed my assignment and gave his valuable Feedback.

Feedback from Supervisor:

	Feedback	Response
1	Research Question is not clear	Reframed research question
2	Data Analysis is not mentioned	Mentioned Data Analysis
3	Publication and Printer material	Mentioned about Publication and Printer material
4	Consider various other fields	Considered various other fields such as forensic science to identify criminals.
5	Researcher can use this knowledge for?	Mentioned this concept.