

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

## Understanding Research

Vinay Huchanahalli Nagaraju – N10180893

Supervisor : Dr. Patrick Delaney

## **Content:**

1. Problem Statement -----	02
2. Research Questions -----	02
3. Question Comparison -----	06
4. References -----	06
5. Reflective Statement -----	07
6. Appendix -----	08

## **Abbreviations:**

**CNN – Convolutional Neural Network**

**ANN – Artificial Neural Network**

**RAM – Random Access Memory**

**UV rays - Ultraviolet rays**

**QUT – Queensland University of Technology**

**CPU – Central Processing Unit**

**GPU – Graphical Processing Unit**

**RNN – Recurrent Neural Network**

**LR – Linear Regression**

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

## **1. Problem Statement:**

Skin cancer is the development of abnormal cells in the body. Skin cancer is mainly caused by sun's rays inflicting the human skin, causing abnormal growth in skin cells. Greater than 90% of cases caused by exposure to the ultraviolet (UV) radiation from the Sun. There are 3 types of skin cancer. Melanoma is the most dangerous type of skin cancer which is caused by the UV rays from the Sun. In Australia, 12,500 new cases of melanoma are reported each year, in which more than 1500 die from this disease. Melanoma makes up only 2.3% of all skin cancer but it is responsible for 75% of skin cancer deaths. 1 in 17 Australians will be diagnose with Melanoma before age 85 (Sinclair, 2013)<sup>i</sup>. Melanoma can be cured if it detects in early stage. 1 in 20 American adults have been misdiagnosed in the past and over half of these is harmful. Convolutional Neural Network (CNN) will help in this case. CNN will help to classify skin cancer based on the marks on skin. Patient can take the image and upload in computer and CNN algorithm will classify whether it is skin cancer or not. There are 3 limitations in CNN. One such limitation is lack of data. CNN will learn classification based on the previous cases. It needs more data to classify the cancer. But, there is not enough data is available due to improper data collection (Sen Sharma, Ahmed Mridul, M & Saiful Islam, 2018)<sup>ii</sup>. Hospitals don't have any structural procedures to collect the images of the cancer from patients and provide it to skin cancer research centres. Second limitation is Computational and space complexity. Convolutional Neural Network is a powerful and complex Algorithm<sup>iii</sup>. It needs Graphical Processing Unit (GPU) and High Random Access Memory (RAM) to compute the data. RAM is the primary memory in any computer system (Ardakani & Condo, 2018). Ordinary/Office computers work on Central Processing Unit (CPU). Hence, more money is needed to implement the GPU and RAM in normal or office computers. Third limitation is the quality of the images. If the pixel of the image is not good then CNN will not predict the presence of skin cancer accurately. Hence, it is mandatory to have good camera to take the better images for Convolutional Neural Network. Filter or Kernel plays the important role in the detection of the edge in CNN. If the detection of the edge is not accurate then it leads to the misclassification of skin cancer (Shin & Yun, 2019)<sup>iv</sup>. Hence, developing the kernel is very important in order to increase the performance of CNN.

## **2. Research Questions:**

### **2.1. How to improve the data collection process in a systematic way for Convolutional Neural Network in Australian Hospitals to improve the performance of skin cancer detection.**

Data collection is one of the main problem in Image classification. The Intuition of wrong or empty data leads to the misclassification of skin cancer. Moreover, this wrong prediction will cost lives. Australian Hospitals don't have systematic process to collect the data from the patients. They diagnosis the patient and predict whether he is having cancer or not. It is necessary to implement some rules and regulations to collect the Data and ask them to store

in those data in one place (Sen Sharma, Ahmed Mridul, M & Saiful Islam, 2018). Machine Learning Engineer can use this data to train the Convolutional Neural Network. Due to the more data, the model or Algorithm will perform well.

By exploring this question, it is easy to understand the different rules and procedures to collect the data from different regions of the world. In Australia, the pattern of skin cancer may be bit different than rest of the world. So, we will get the clarification about this, if the data is available in one place. This will help to improve the accuracy of the skin cancer prediction using CNN. This will help to make all the researcher in different part of the world to work together on skin cancer disease.

Systematic way of approach is needed to gather the Data from Australian Hospitals. Since, Australia is having the highest number of skin cancer patient in the world, it will benefit to those patients. Survey Method is the best method to improve the data collection process and Factor Analysis helps to evaluate the performance of new data collection process. Clinicians will be conducting the survey and store it in a local systems or computers. This computer must synchronised with Skin Cancer Research Centre (SCRC) in Australia. All data from different hospitals are collected on each and every day in skin cancer research centre. People who are present in the research centre are responsible to manage and maintain those data. Currently, all patient data is distributed across different Hospitals. Hence, this process will help to collect those data from one source.

Clinicians must conduct the surveys among the patients who visit the hospital and collect the data from them. Data must include all important features such as pattern of mark on skin, size of the mark and other important information. This will improve the quality of the Data (Sen Sharma, Ahmed Mridul, M & Saiful Islam, 2018). By collecting the data from survey methods, Researcher can integrate those Data from SCRC and do the Data cleaning process. It will improve the quality of the collected data and improves the performance of the Convolutional Neural Network Model.

## **2.2. What enhancements will improve image classification performance in Convolutional Neural Network model to detect skin cancer in Australia?**

Image classification plays an important role in the detection of skin cancer. If the image classification is accurate then the performance of the CNN model improves. High accuracy of the model leads to the accurate detection of skin cancer. Since, there are more number of skin cancer patients are present in Australia, accurate prediction of skin cancer will help to benefit to them. There are many cancer research institutes who are working on improving the image classification. Queensland University of Technology is also investing more resource on image classification to detect the cancer.

By exploring this question, we will get the factors which will improve the image classification. Currently, kernel/filter is the best method to detect the edges in the images. It is not 100% accurate. Hence, the solution from this question will leads to the invention of another methodology to detect the edges of the images. It will help to improve the classification process. Since, image plays a vital role in the cancer detection (Shin & Yun, 2019). Accurate classification of images will benefit to the patients.

Quantitative method is more suitable to analyse the performance of the CNN model. Statistical modelling such as hypothesis testing is helpful in evaluating the performance. Hypothesis testing consists of 2 parameters. Null Hypothesis and Alternative Hypothesis. It

is necessary to disprove the null hypothesis. Let us consider null hypothesis as  $H_0$  and Alternate Hypothesis as  $H_a$ .

$H_0$ : Performance of CNN will not improve.

$H_a$ : if we improve the kernel or filters then performance of CNN will increase.

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

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

Alpha and beta values are misclassified errors generated during the Hypothesis testing.

It is necessary to make  $H_0$  statement as false. Hypothesis testing will help to determine whether the improvised kernel help to improve the image classification in CNN or not.

### 2.3. What underlying factors affect the performance of the Convolutional Neural Network Model?

The main factors which are affecting the performance of the CNN is the quality of the image and the computational complexity to run the CNN in normal computers. CNN is a powerful machine learning model. Hence, it takes more RAM and GPU power to do the computation (Hung, Tang, Wang, & Chen, 2018)<sup>v</sup>. This will give the brief idea about performance and accurate classification of the model. If we understand the factors behind the performance of the Convolutional Neural Network then we can improve the performance of the model. Hence, we can improve the cancer detection in Australia.

By exploring this question, we can explore the ideas which are affecting the performance of the model. It will help to minimize the time and space complexity to run the model. Sometimes, it will also help to produce the new machine learning model like Convolutional Neural Network. The new model may take less computational speed. It may have less time and space complexity and it may perform well in the blur image as well. If we can understand what factors in this context then we can introduce a New Algorithm or Model to detect the cancer that will enhance the performance of Convolutional Neural Network.

Quantitative method is more suitable to find the solution for this question. It is necessary to group all the factors which are impacting the performance of the Convolutional Neural Network. Statistical Analysis such as Logistic Regression (LR) will help to find whether the new model is performing better than the old CNN model. Logistic regression will help to classify the factors. The factors with positive impact on the model will be classified as 1 and factors with negative impact will be classified as 0. Logistic Regression use sigmoid function to classify the factors.

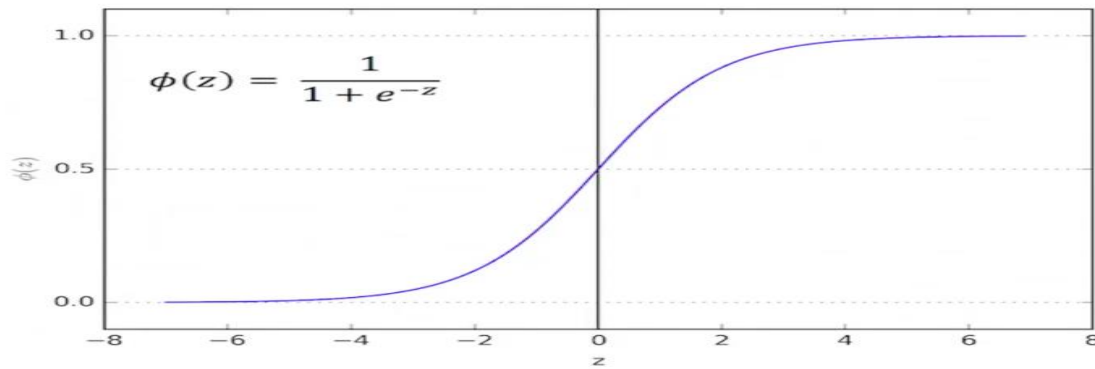


Image2: Source: <https://towardsdatascience.com/logistic-regression-a-simplified-approach-using-python-c4bc81a87c31>

In Logistic Regression, threshold value will be 0.5. If the accuracy of the factor lies within 0.5 then it will be classified as 0 else it can be classified as 1. Accuracy of the factors can be calculate using Precision and Recall.

		Predicted class	
		<i>P</i>	<i>N</i>
Actual Class	<i>P</i>	True Positives (TP)	False Negatives (FN)
	<i>N</i>	False Positives (FP)	True Negatives (TN)

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

Image3: Source: [http://rasbt.github.io/mlxtend/user\\_guide/evaluate/confusion\\_matrix/](http://rasbt.github.io/mlxtend/user_guide/evaluate/confusion_matrix/)

If the factors which are having High Precision and Recall is having high accuracy. The factors with high accuracy will impacting positively on the model. Which help to increase the accuracy of CNN to predict the cancer. Model will work well if all the negative factors are removed.

### 3. Question Comparison

First question is related to Data Collection procedure, which is a general problem and it is not specific to the performance of CNN model. Hence, it is not a vital question that needs to be answered. Data Collection, pre-processing, and cleaning is the common procedure that needs to complete before contracting any model. It will help to improve the performance of the Model. Which is not specific to detect the cancer. Set of rules and procedures will not vary much when compare to Australia and other countries.

Second question is most important and more relevant to the main research Question. Because, the main intention of this Literature review is to increase the performance of the CNN model. Image classification will help to improve the accuracy of the image detection which inters improves the performance of the Convolutional Neural network (CNN) model. Hence, this question will help to find the solution to the main problem. The main constrain of this question is to find the filter or kernel that detect the edges. Current filter or kernel is not that much accurate (Ardakani & Condo, 2018). Research is in progress to find the improvised kernel. Hence, it needs more time to find the accurate filters.

Third Question is pointing to find the underlying factors that impacting the performance of CNN model. It also an important question. But, this question is pointing to the general factors such as time and space constrains. This question is not that much specific to the main question. But, answering this question will improve the performance of CNN model inters increase the cancer classification.

Overall, second Question is more specific to this literature review. Skin cancer can be classified based on the images of the marks on the skin. If the image classification is not accurate then it is not even possible to classify the cancer. It will also leads to further research on kernel and filter improvisation. It may also leads to the invention of new powerful algorithm like Convolutional Neural Network. Hence, it is important to answer the second Question.

### 4. Reference:

---

<sup>i</sup> Sinclair, R. (2013). Nonmelanoma skin cancer in Australia. *British Journal of Dermatology*, 168(1), 1–2. <https://doi.org/10.1111/bjd.12167>

<sup>ii</sup> Sen Sharma, A., Ahmed Mridul, M., Jannat, M., & Saiful Islam, M. (2018). A Deep CNN Model for Student Learning Pedagogy Detection Data Collection Using OCR. In *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)* (pp. 1–6).

<sup>iii</sup> Ardakani, A., Condo, C., Ahmadi, M., & Gross, W. (2018). An Architecture to Accelerate Convolution in Deep Neural Networks.

<sup>iv</sup> Shin, C., & Yun, J. (2019). Deep Rotating Kernel Convolution Neural Network. In *2019 Third IEEE International Conference on Robotic Computing (IRC)* (pp. 441–442).

<sup>v</sup> Hung, C., Lin, Y., Tang, C., Wang, C., & Chen, M. (2018). Performance of Convolution Neural Network based on Multiple GPUs with Different Data Communication Models

<sup>v</sup> Hung, C., Lin, Y., Tang, C., Wang, C., & Chen, M. (2018). Performance of Convolution Neural Network based on Multiple GPUs with Different Data Communication Models

---

## 5. Reflective Statement:

**Student : Vinay Huchanahalli Nagaraju**

**ID: N10180893**

**Supervisor : Dr. Patrick Delaney**

<b>Week</b>	<b>Supervisor</b>	<b>Student</b>
6	Patrick taught me how to raise the research question. He took the real world example such as Pacemaker technology. He raised 4-5 research questions and explained the importance of raising the research questions. He explained the reason of solving the research question.	I raised 4 research question which is related to my research domain. I understood the importance of raising the research question. Those 4 research questions are distinct and they addressed the different problems and limitations.
7	He took those research questions and explained the reasons to solve them. He also explained the following things for each research question 1. How it contributes to knowledge? 2. How it solves the overall problem? 3. What method, process or technique will you use?	I explored the new knowledge, limitations, advantage and disadvantage for each questions using different sources such as Literature reviews, articles, case studies.
8	He mainly focused on the research methodology. He explained the various types of methodologies such as Qualitative and Quantitate methodology to answer the research questions.	I understood the various methodologies such as Qualitative, Quantitative and mixed methodologies to answer the research questions.
9	He discussed about the marking criteria for assignment 2. We went through each section of marking criteria sheet and explained what he is expecting from the students.	I understood the marking criteria. At the end of the tutorial, I got feedback from Patrick for my drafted assignment. I understood my mistakes in this assignment.



---

## 6. Appendix : Feedback from Supervisor

	Feedback	Response
1	Add the relevant reference in methodology.	I added the relevant source or methodology.
2	Try to avoid "How..?" or "In order to" while constructing the questions.	Avoided How...? Type of question and deleted "in Order to" in the question.
3	It is not good/appropriate to add mixed methodologies for this type of question.	Changed the methodology to Quantitative because my focus is to improve the performance of CNN.
4	Write <u>precise</u> statements (e.g. "Skin cancer is the development...", "Melanoma can be cured if detected early...").	Changed the sentences and made that more precise.
5	Question 1 Methodology is not clear.	Changed the methodology and made it clearer.
6	Change the Question 2 as "What enhancements will improve image classification performance in...."	Changed the Question 2.
7	Q3 - watch out for proofreading, should be "affect"	Corrected the grammatical error changed the sentence from "affecting" to "affect".
8	The method here seems more like a two-phase approach - a qualitative strategy to develop the classification, then simulation data using a quantitative strategy...? I'm not sure, this part possibly needs further clarification/consideration, just make sure the process is clear.	Changed the methodology to Quantitative approach and made it clearer. I applied the Logistic Regression for the classification of the factors.