

Skin cancer detection

By Nicolai Plambeck Hansen



April 18, 2022

University of technology Sydney (UTS)

Institute of Data

Table of Contents

[Introduction 3](#_Toc101207721)

[Problem statement 3](#_Toc101207722)

[What is the problem that needs to be solved? 3](#_Toc101207723)

[Why is this problem valuable to address? 3](#_Toc101207724)

[What is the current state? 3](#_Toc101207725)

[What is the desired state? 3](#_Toc101207726)

[Has this problem been addressed by other research projects? 3](#_Toc101207727)

[Industry / Domain 4](#_Toc101207728)

[Industry 4](#_Toc101207729)

[Market Drivers 4](#_Toc101207730)

[Top players 4](#_Toc101207731)

[Market Value 4](#_Toc101207732)

[Business question 5](#_Toc101207733)

[Stakeholders 5](#_Toc101207734)

[Who are the stake holders? 5](#_Toc101207735)

[Outcome expectations 5](#_Toc101207736)

[Data question 5](#_Toc101207737)

[Data Process 6](#_Toc101207738)

[Data acquisition 6](#_Toc101207739)

[EDA 6](#_Toc101207740)

[Data Pre-processing 6](#_Toc101207741)

[Modelling 7](#_Toc101207742)

[Model function used 7](#_Toc101207743)

[Best model 7](#_Toc101207744)

[Outcome 7](#_Toc101207745)

[Implementation 7](#_Toc101207746)

[Data Answer 8](#_Toc101207747)

[Business answer 8](#_Toc101207748)

[Recommendation to stakeholders 8](#_Toc101207749)

[End-to-end solution 8](#_Toc101207750)

[Reference List 9](#_Toc101207751)

[Documentation 9](#_Toc101207752)

[Models 9](#_Toc101207753)

[Data 9](#_Toc101207754)

# Introduction

Australia has one of the highest rates of skin cancer in the world. Is it expected that two out of three Australians will be diagnosed with skin cancer before they turn 70 and around 2000 will die from skin cancer each year. The sooner skin cancer is detected the better the chances are to not only survive but also avoid long hospitalisation, surgery or disfiguration.

Australia’s rate of skin cancer is roughly two to three times higher than that of the UK, USA or Canada.

# Problem statement

## What is the problem that needs to be solved?

Not everyone has easy access to a dermatologist/doctor or a way to access if a mole is malignant skin cancer.

We need to ensure that everyone, especially those in remote areas, can get easy access to means of diagnosing/accessing if a mole is malignant skin cancer.

## Why is this problem valuable to address?

Malignant cancer cells get classified in stages from 0 to 4. The treatment of skin cancer gets progressively more expensive as the cancer cells moved from stage 0 to 4.

In 2018/2019 the Australian health system spend more than $1.68 billion on skin cancer. Of this number $357 million was attributed to melanoma. It is the most expensive cancer type to treat.

A study published in the Archives of Dermatology found that the costs for skin cancer treatment ranges from USD $1,732for stage I disease to USD $56,059for stage IV disease.

## What is the current state?

Currently there’s no formal screening program for skin cancer in Australia. It relies on people becoming familiar with their skin and if they notice any changes to contact a doctor.

## What is the desired state?

The desired state is to have an algorithm that can fairly reliably highlight if a mole is a cause for concern and should be investigated further.

## Has this problem been addressed by other research projects?

Recently MIT has developed a skin cancer detection AI. It achieved an accuracy of over 90.3% in distinguishing suspicious lesions from non-suspicious ones without an individual lesion image.

# Industry / Domain

Industry

For this project we will be focusing on the health care industry as large investments are being made towards the research and treatment for skin cancer.

## Market Drivers

Skin cancer is diagnosed by a physical examination and in some cases where there might be a concern, a biopsy.

A biopsy is a simple procedure where part or all of the mole is removed and sent to a laboratory for further testing.  
  
Major market drivers are: the increasing incidence of melanoma and non-melanoma skin cancer, technological advancements, imaging techniques that offer confirmatory diagnostic tests and product differentiation to drive growth and prevent the threat of substitutes..

## Top players

* Alma Lasers
* Agilent Technologies Inc
* Biolitec Ag
* Bruker Corp
* Ellipse A/S
* GE Healthcare
* Leica Microsystems
* Michelson Diagnostics
* Syneron Medical
* Toshiba Medical Systems

## Market Value

According to Fortune Business Insight, the global skin cancer treatment market was valued at 8.19 billion in 2019 and projected to reach 14.55 billion by 2027.

# Business question

**Can we develop a model which will reduce the expense and time spent on biopsy, treatment and diagnoses of skin cancer?**

# Stakeholders

## Who are the stake holders?

* Government health departments
* Hospital
* Clinic
* Laboratory
* General public

## Outcome expectations

Stakeholders would expect to see a reduction in spend on the treatment of skin cancers as well as reduced hours spent on analysing samples..

# Data question

**Can we determine from a picture of a mole if it is comprised of benign or malignant skin cells?**

To help answer this question we will need a large number of mole images. These images need to be of different moles that are confirmed to be either malignant or benign. Preferably the images will come in varying light and posture. If this is not the case, we can artificially implement these conditions using Keras ImageDataGenerator which could also help us save memory when running our models.

# Data Process

## Data acquisition

The data for my project was sourced from Kaggle. (<https://www.kaggle.com/datasets/fanconic/skin-cancer-malignant-vs-benign>)

The data consists of 3297 images of moles

The images are sorted into a test and a training folder. Each of these folders are further spilt into a benign and malignant folder.

The dataset is balanced between benign and malignant.

Each picture is 224 x 224 and in RGB format.

The data and stems from The International Skin Imaging Collaboration (ISIC, <http://www.isic-archive.com>)

ISIC runs yearly competitions for skin cancer predictions. The latest database consists of more than 20000 pictures.

## EDA

Various images from each set are inspected in the EDA. To make the distinction between higher and lower resolution images both are displayed showing the same picture. The pictures displayed during the EDA will vary each time the code is run.

A histogram is used to display the quantity of each category so we can highlight inspect if it is a balanced dataset we are working on

## Data Pre-processing

Images were resized to either 32 x 32 or 64 x 64 before being loaded into the model. This is done to see if there is going to be a significant difference in training time and effect.

I have also normalized all pixel values to range between 0 and 1 instead of the standard 0 and 255. This is achieved by dividing by 255.

The images were imported using glob and the targeted values were categorised using the tensorflows build in utility.

## Modelling

For prediction and evaluating the images these models have been used:

* Conventional Neural Network
* ResNet50
* VGG16

Each of these have been used two times to see test if there would be a difference based on resolution.

### Model function used

Functions to utilize early stopping training of the model has been implemented when accuracy stops improving.

Functions to slow learning rate has been implemented to help improve the model and achieve it maximum potential.

## Best model

From our test of different models, the one that performed best was the VGG16.

The model was trained using pictures resized to 32 x 32 RGB

The model was stopped by our early stopping function and the scores from epoc 50 were picked as the optimal weights for our network.

It achieved an accuracy of ~85% with a loss of ~0.33 on our test data.

## Outcome

From the tests it appears that a lower resolution in general yield a better result. This could of course change once I start investigating the effect of using high resolution pictures of more than 1024 x1024, as you could speculate that the higher resolution would help models to better pick out distinct features of the pictures.

The advantage of using lower resolution pictures is that we can train models a lot faster giving us time for training the model on more data. It will also be able to run on smaller systems and not require dedicated / high end CPU or GPU.

## Implementation

Before this model can be implemented in a wider setting more testing and tuning is required.

It could be very interesting to fit the model with background information on the patient. Information such as age, location of mole, ethnicity, geographical location of the patient etc. could potentially improve the model dramatically.

It could be interesting to implement this model as an app or a website where people would be able to use their smartphones to take a picture of a mole of concern and have the model evaluate it.

My biggest implementation is concern is giving someone a benign prediction for their mole of concern, when it is in fact malignant as this may cause someone to not have that mole checked by a professional and they do not get treatment.

# Data Answer

We can conclude that it is possible to give a good estimate whether or not a mole is cancerous based on just an image. But as a wrongful prediction can have devastating effects here I would not recommend to push this option for the general public just yet.

We have with fairly limited information achieved an accuracy score of 85% which I would consider reasonable from the expected data available.

# Business answer

We can develop a model which can give an indication of the risk of a mole being cancerous but we can’t yet reduce the time and effort spent on doing manual investigations of moles. This might support doctors in spotting moles of concerns early on and reduce cost of treatment that way.

So we can develop a model which can give and indication but not a conclusion.

# Recommendation to stakeholders

I would not put this model into production yet as it, unfortunately, still predicts some malignant moles as benign which is our biggest issue.

The model could potentially be used as a supplement to medical professionals who would also be able to manually access the state of the mole. The model could help speed up predictions.

# End-to-end solution

We need to collect more data especially on the background of the mole. This can potentially be used to further improve the model and get it to a point where the general public would be able to use it as a reliable estimate.

Once we get a model with an accuracy rate of above 99% I would suggest releasing it to the general public as a reliable alternative. This could hugely benefit the medical and rural communities.

# Reference List

## Documentation

* <https://www.datamintelligence.com/research-report/skin-cancer-market/>
* <https://www.fortunebusinessinsights.com/skin-cancer-treatment-market-102806>
* <https://health.costhelper.com/skin-cancer.html>
* <https://actcancer.org/prevention/sunsmart/skin-cancer-in-australia/>
* <https://analyticsindiamag.com/mit-researchers-develop-an-ai-tool-to-detect-skin-cancer/>
* https://www.marketwatch.com/press-release/skin-cancer-diagnostics-market-share-2022-comprehensive-analysis-growth-prospects-top-key-players-size-estimation-product-scope-and-forecast-to-2025-2022-04-13

## Models

* Keras.io
* <https://www.kaggle.com/code/sid321axn/step-wise-approach-cnn-model-77-0344-accuracy>
* <https://www.kaggle.com/code/fanconic/starter-skin-cancer-malignant-vs-benign>
* <https://www.geeksforgeeks.org/visualizing-representations-of-outputs-activations-of-each-cnn-layer/?ref=rp>
* <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8684510/>

## Data

* https://www.kaggle.com/datasets/fanconic/skin-cancer-malignant-vs-benign