A Project Report on

**“CANCER DETECTION USING CNN”**

**Submitted in partial fulfillment of the requirements of Post Graduation Diploma in Big Data Analytics**

**(PG-DBDA)**

**SUBMITTED BY**

**SEPTEMBER 2022**

**RAHUL VASANT WAKALKAR (220910125021)**

**AARUSHI SINHA (220910125010)**

**NISHIKANT VISHNU NEMADE (220910125018)**

**PANKAJ KUMAR (220910125029)**

**GUIDEDED BY**

**Dr. ZEESHAN AHMED KHAN**

**Faculty, CDAC Delhi**

**Center for Development of Advanced Computing Delhi**

**CERTIFICATE**

This is to certify that the Report work entitled

**“CANCER DETECTION USING CNN”**

Has been duly completed by the following students under the guidance, in a satisfactory manner as a partial fulfillment of the requirement for the award of the Post Graduation Diploma in Big Data Analytics.

SUBMITTED BY

SEPTEMBER 2022

Rahul Vasant Wakalkar (220910125021)

Aarushi Sinha (220910125010)

Nishikant Vishnu Nemade (220910125018)

Pankaj Kumar (220910125029)

**Mr. Ankit Khurana Dr. Zeeshan Ahmed Khan**

**Principal Technical Officer Project Guide**

**DECLARATION**

I declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

---------------------.

Rahul Vasant Wakalkar (220910125021)

---------------------.

Aarushi Sinha (220910125010)

---------------------.

Nishikant Vishnu Nemade (220910125018)

---------------------.

Pankaj Kumar (220910125029)

DATE:12/03/2023

**ACKNOWLEDGEMENT**

We are heartily thankful to our guide **Dr. Zeeshan Ahmed Khan**, for his guidance patience and support. We consider ourself very fortunate for being able to work with very considerate and encouraging professor like him. Without his offering to complete these study work, we should not finish our project. It is also our duty to record thankfulness to our whole CDAC department for their help in needs.

Our special thanks to our parents and all of friends for help us exchanging any ideas and give the enjoyable study environment. At last, we special gratify to almighty God for blessing us with the hidden power to completing this study work.

**PROJECT TEAM**

Rahul Vasant Wakalkar (220910125021)

Aarushi Sinha (220910125010)

Nishikant Vishnu Nemade (220910125018)

Pankaj Kumar (220910125029)

**TABLE OF CONTENTS**

**1.ABSTRACT……………………………………………………………… 6**

**2.INTRODUCTION………………………………………………………… 7**

**3.TECHNOLOGIES………………………………………………………… 8**

**4.HARDWARE REQUEREMENTS FOR IMPLEMENTATION AND RESULT ANALYSIS……………………………………………………….. 11**

**5.SOFTWARE REQUIREMENTS………………………………………… 11**

**6.DATA SET DESCRIPTION……………………………………………… 12**

**7.DISCRIPTION OF LIBRARIES USED…………………………………. 13**

**8.TOOLS USED……………………………………………………………... 15**

**9.SIMULATION RESULT…………………………………………………. 16**

**10.RESULTS………………………………………………………………… 26**

**11.CONCLUSION…………………………………………………………... 26**

**12.FUTURE WORK………………………………………………………… 27**

**13.REFERENCES………………………………………………………….... 27**

1. **ABSTRACT**

Computer vision is a branch of computer science that focuses on teaching computers to see, recognize, and interpret data in the same way that humans do, and then produce appropriate results. It's like educating a machine human intelligence and impulses. It comprises techniques for receiving, processing, analyzing, and comprehending video and image data. In today's world we are fortunate enough to have advanced technology which allows the medical field to provide better patient care. Cancer is a highly researched area because there are so many people who suffer and/or die of cancerous disease. This study aims to improve cancer detection in lymph nodes by using computer vision machine learning techniques. We will examine the data to get a better understanding of it. Then, we will run multiple convolutional neural network models with the intent to be able to classify cancerous (1) and non-cancerous cells (0). With improved and faster cancer detection, patients will be able to receive life-saving treatments faster.

1. **INTRODUCTION**

Information technology IT plays a huge role in diagnosis of diseases. Information technology improvement offers advancement in health care in terms of diagnosis, management, and support of the diseases. The physicians also get aids in diagnostic process due to improved decision support systems incorporated with evidence base medicines. Many well studies have been established in cancer detection and diagnosis by using information technology. Some of them applied the machine knowledge methods for example Naive Bayes, Random Forest, Agent technology and Neural Network (in 2020). Whereas, the best results in cancer detection and diagnosis have been observed with Deep learning techniques such as Convolution Neural Network (CNN).

Cancer is the second leading cause of death worldwide, an average of one in six deaths is due to cancer. Considerable research efforts have been devoted to cancer diagnosis and treatment techniques to lessen its impact on human health. This study aims to improve cancer detection in lymph nodes by using computer vision machine learning techniques. We will examine the data to get a better understanding of it. Then, we will run multiple convolutional neural network models with the intent to be able to classify cancerous (1) and non-cancerous cells (0). With improved and faster cancer detection, patients will be able to receive life-saving treatments faster.

1. **TECHNOLOGIES**

Hierarchical learning or deep structured learning are terms used to describe deep learning. Artificial neural networks are a subsection of deep learning that is interested in algorithms inspired by brain anatomy and function. Deep learning models are based on biological nervous system information processing and communication patterns. Deep learning is a type of machine learning technique that extracts and transforms features using a cascade of numerous layers of nonlinear processing units. The output from the preceding layer is used as input for each subsequent layer. Learning takes place across a hierarchy of layers of representations that correspond to different levels of abstraction.

CNN stands for Convolutional Neural Network and it is a feed forward Neural Network that can classify images. It just considers current input. Convolution, ReLU, Pooling CNN's four levels are, and Fully Connected Layers. Each layer has its unique role, which includes feature extraction and the discovery of hidden patterns.

**Flow of CNN**

1.Input Image

The Convolution layer is the most important part of a Convolutional Network because it handles the majority of the computation.

2.Convolution

It consists of ConvNets which are used to match pieces of image and then apply filter.

3.Pooling

The image is shrunk via pooling. Maximum, Minimum, and Average pooling are the three types of pooling.

Because the network must determine whether a new input vector is similar to learned patterns and then create similar output, the backpropagation technique is utilised to discover a local minimum of the error function. The error function's gradient is calculated and utilised to correct the starting weights.

Gradient Descent is a technique for adjusting weight to reduce error and get optimal results. Transfer Learning is a sort of machine learning in which a model developed for one task is transferred to a different task. Given the massive compute and time resources required to develop neural network models for these problems, as well as the massive leaps in skill they provide on related problems, using pre-trained models as the starting point on computer vision and natural language processing tasks is a common approach in deep learning. It is an optimization that allows for faster progress or greater performance while modelling the second task. Two strategies for transferring learning are the Develop Model Approach and the Pre-trained Model Approach. Three common models for transfer learning using picture data are the Oxford VGG Model, Google Inception Model, and Microsoft ResNet Model.

**Image Processing**

The technique of picture processing is an operation applying on an image in order to extract and improve the relevant information. It's a sort of signal processing in which the input is an image and the output is either that image or its characteristics/features. Now a days one of the most quickly evolving technology is Image processing. It is also a critical research field in engineering and computer science. There are three steps involved in image processing which are as follows: Importing an image using image acquisition software; analysing and altering the image; and producing an output that could be an altered image or a report based on image analysis. There are two types of image processing methods employed Analogue and Digital image processing. Hard copies, such as prints and photographs, can benefit from analogue image processing. When applying these visual techniques, image analysts employ a variety of interpretation principles. Computers are used to manipulate digital photographs utilizing digital image processing techniques. When we using digital techniques, we must go through all the three general processes shorts of data Pre-processing, augmentation and presentation, as well as information extraction.

1. **HARDWARE REQUIREMENTS FOR IMPLEMENTATION AND RESULT ANALYSIS**

Memory: 4/8GB

System: Intel i5 2.1 GHz

1TB hard drive

**5. SOFTWARE REQUIREMENTS**

Operating System: Windows 7/8 and above

Domain: Machine Learning

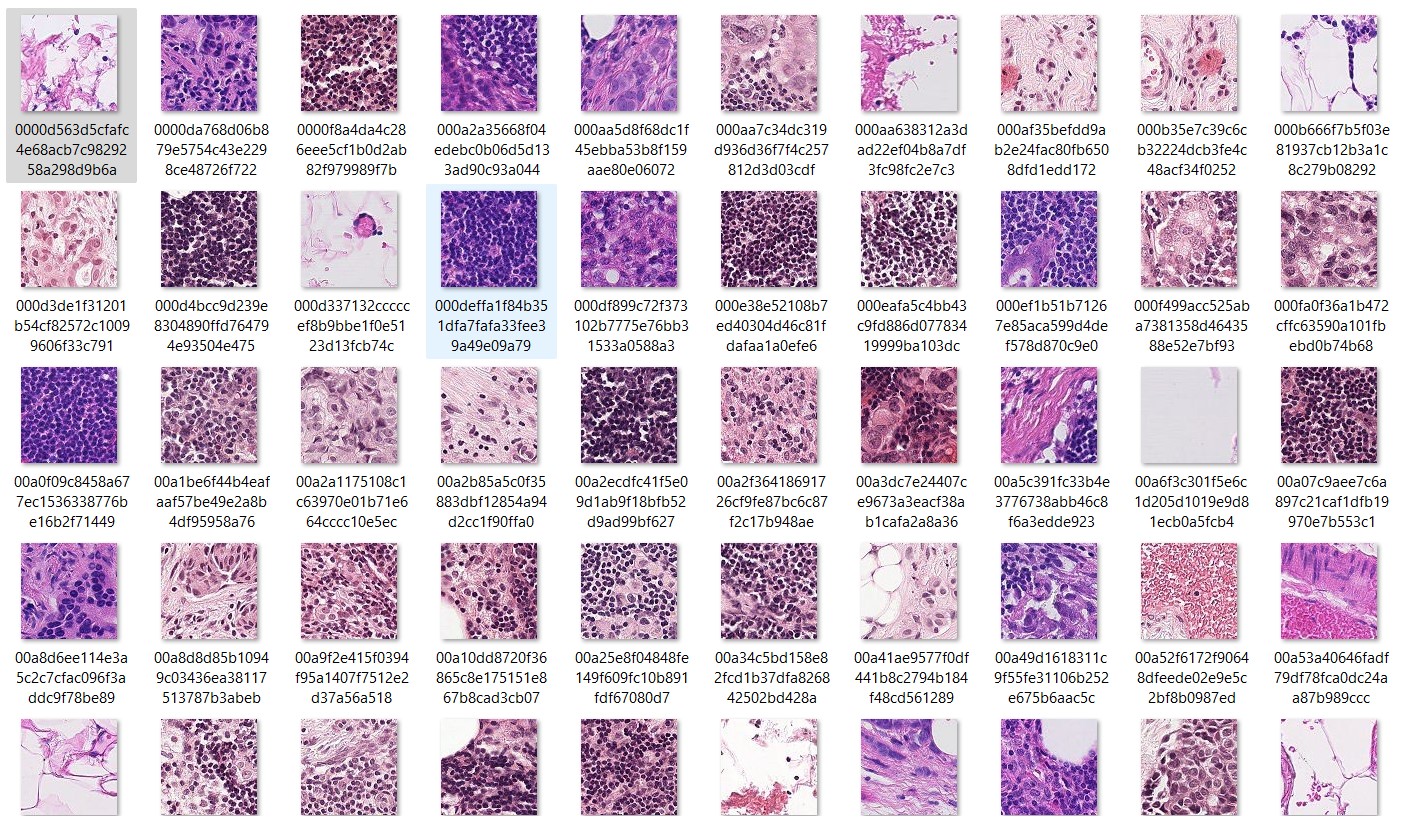
Scripts: Python

Tool: Anaconda Navigator, Jupiter Notebook IDE,

Libraries: Numpy, pandas, matplotlib, seaborn, os, sklearn, keras and TensorFlow.

1. **DATA SET DESCRIPTION**

The dataset contains the histopathological Images, each image is 96px \* 96px. We have two datasets, a training and testing set already split. The training set contains 220,025 unique images and the test set contains about 57,500.

****

**7. DESCRIPTION OF LIBRARIES USED**

* **NUMPY:** NumPy is a simple yet powerful Python library for mathematical and scientific computing and data manipulation. It is a Python library that is open-source.
* **PANDAS:** The Pandas library is used to manipulate and analyse data. It can read and write Excel spreadsheets, CVS files, and do a lot of other things.
* **MATPLOTLIB**: Matplotlib is a graphing library for Python and its numerical mathematics extensions in NumPy.
* **OS:** Python's OS module allows you to use operating system-specific features.
* **TENSORFLOW:** TensorFlow is a free and open-source software framework for dataflow and differentiable programming that may be used to solve a variety of problems. It's a component of neural networks.
* **Seaborn:** Seaborn is a library that uses Matplotlib underneath to plot graphs. It will be used to visualize random distributions.
* **Sklearn:** Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, and clustering and dimensionality reduction via a consistence interface in Python.
* **Keras:** Keras is a python-based open-source high-level neural network API. It makes prototyping simple and quick.

**8. TOOLS USED**

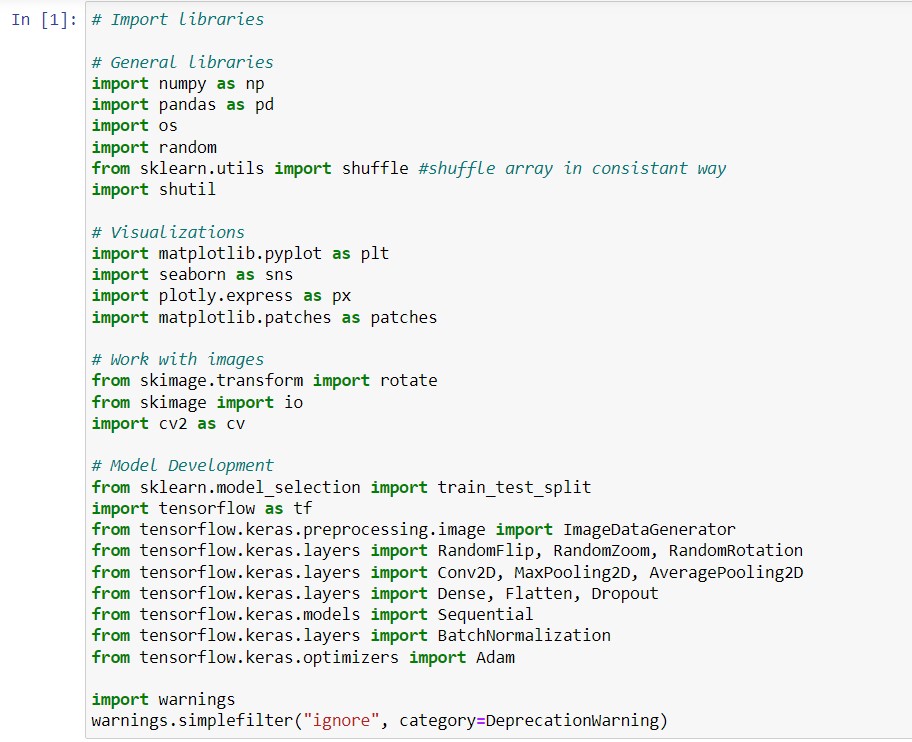
**Anaconda Navigator:**

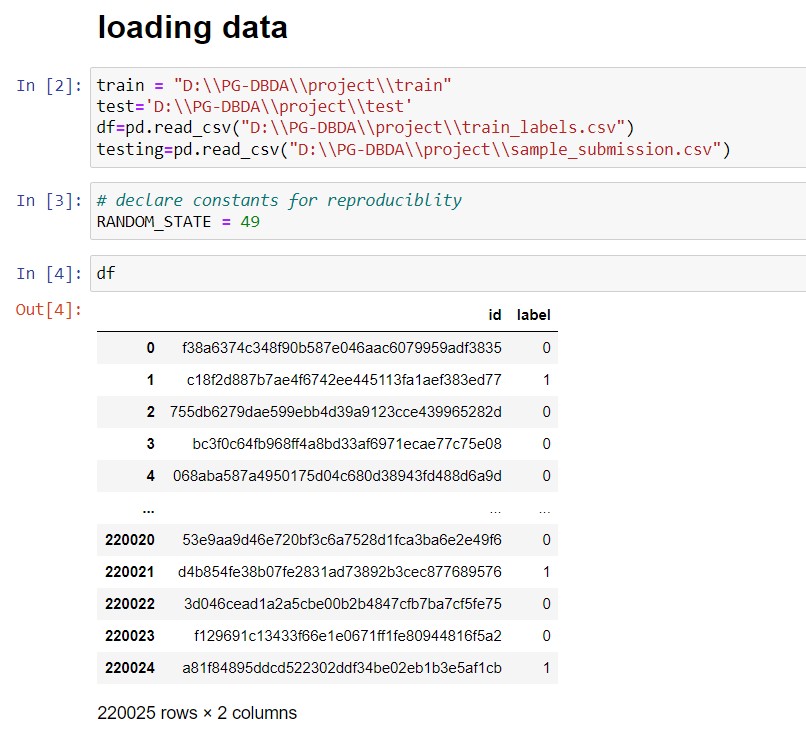
Anaconda is a free and open-source scientific computing distribution that aims to make package management and deployment easier (data science, machine learning applications, large-scale data processing, predictive analytics, and so on).

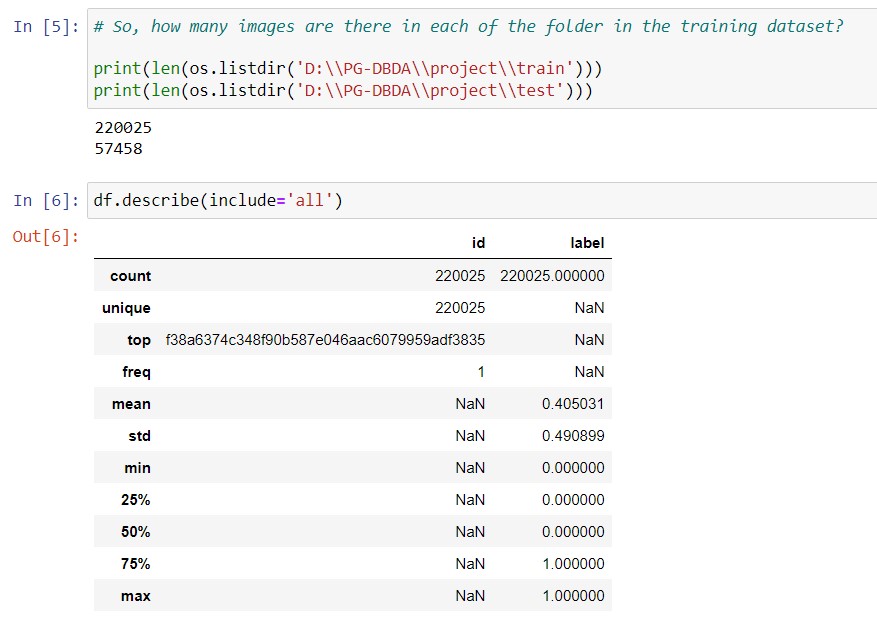
**Jupyter Notebook IDE:**

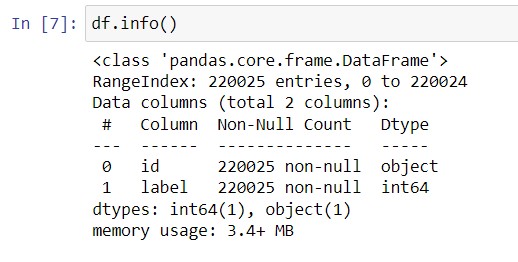
Jupiter Notebook is an open-source and free web tool that lets you create and share documents with live code, equations, visualisations, and narrative text. Data cleansing and transformation, numerical simulation, statistical modelling, data visualisation, and many more features are included in the IDE.

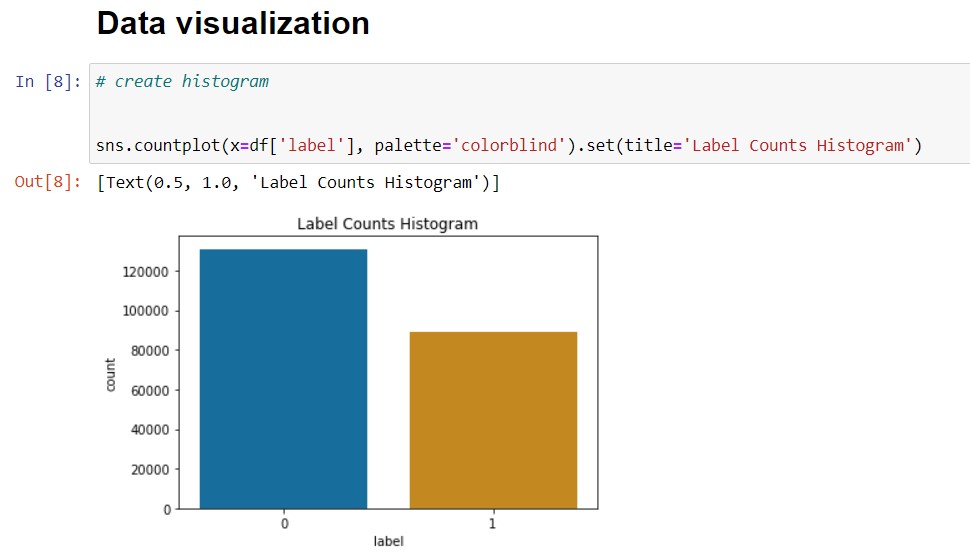
**9. SIMULATION RESULT**

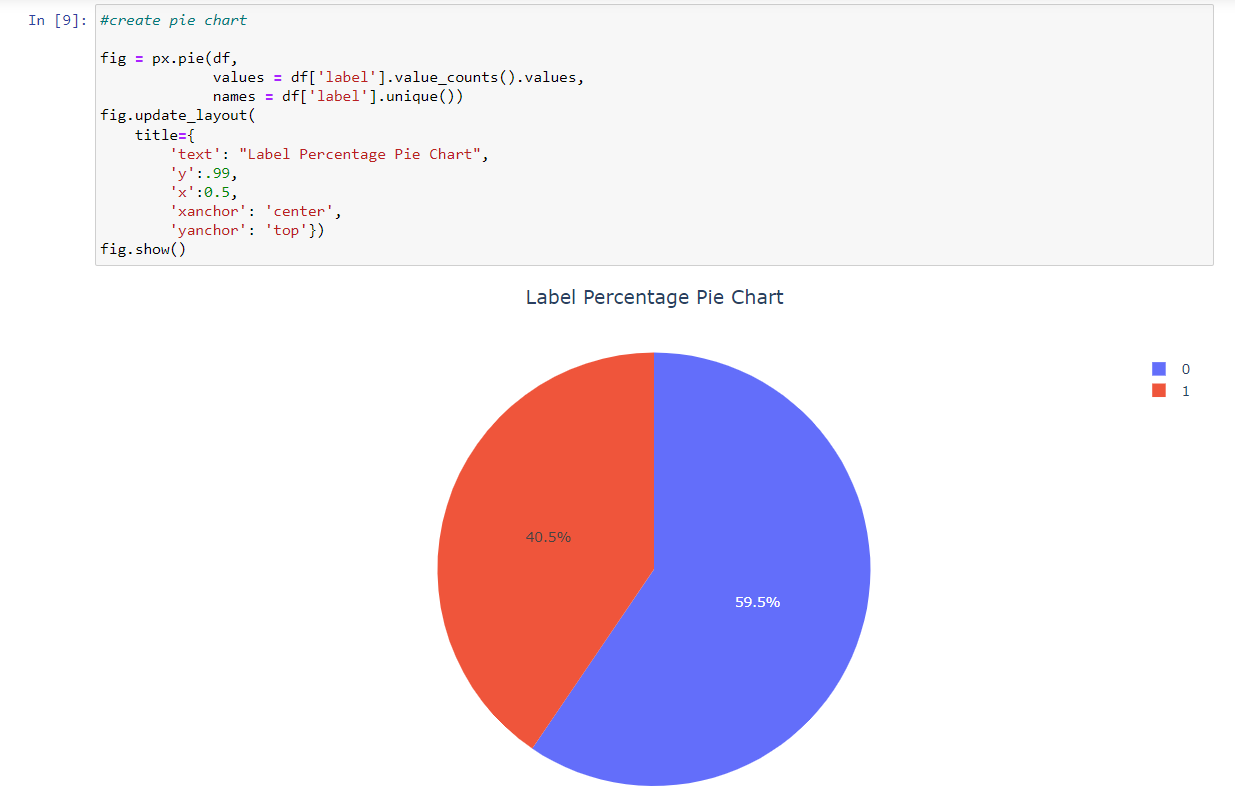
****

****

****

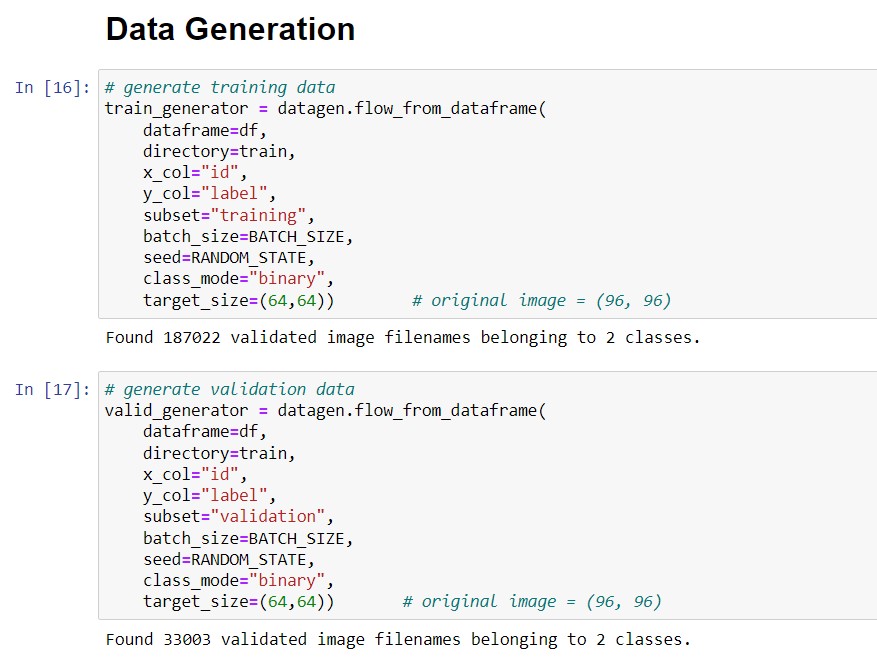
****

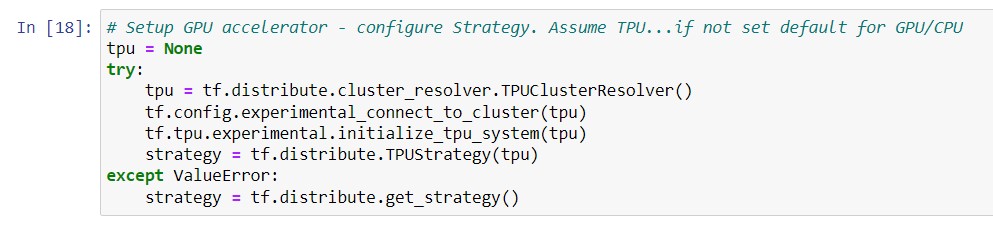
****

****

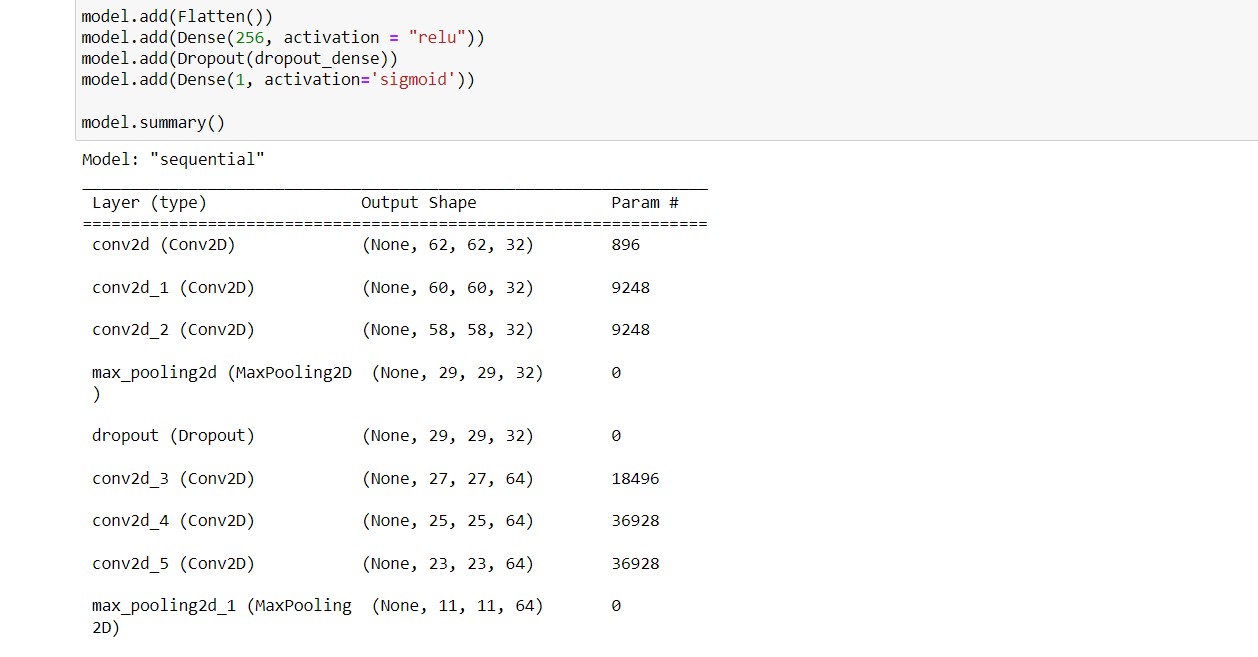
****

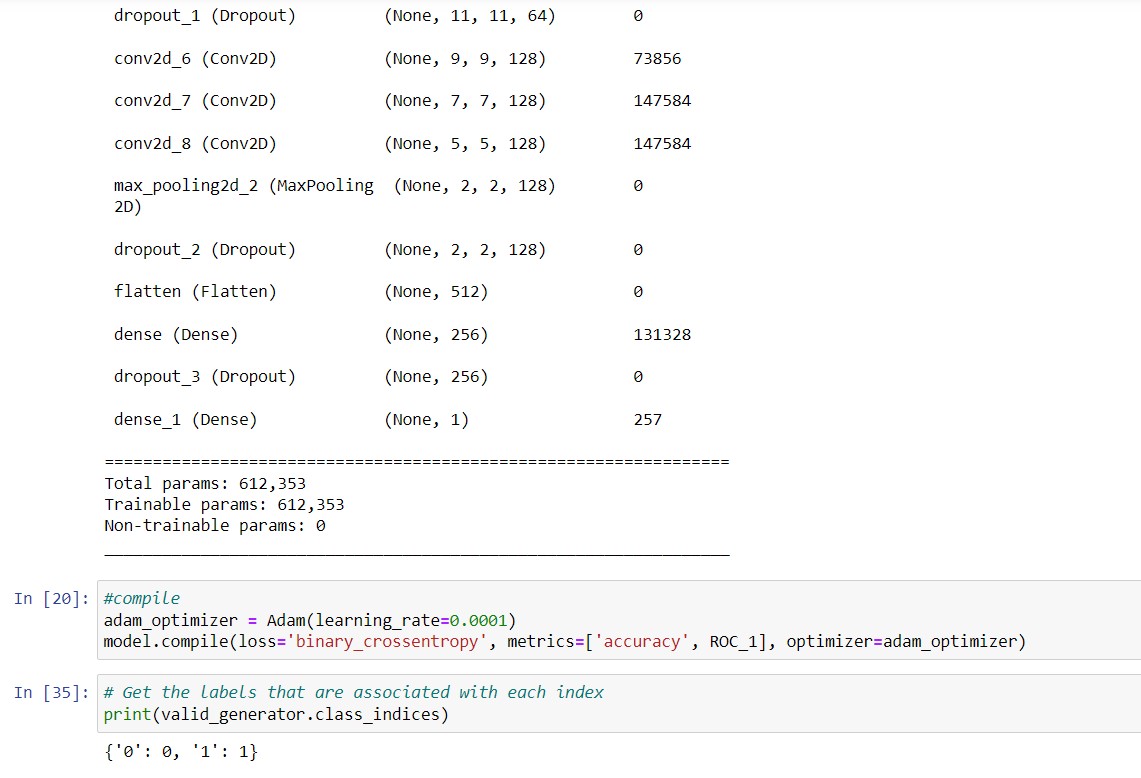
****

****

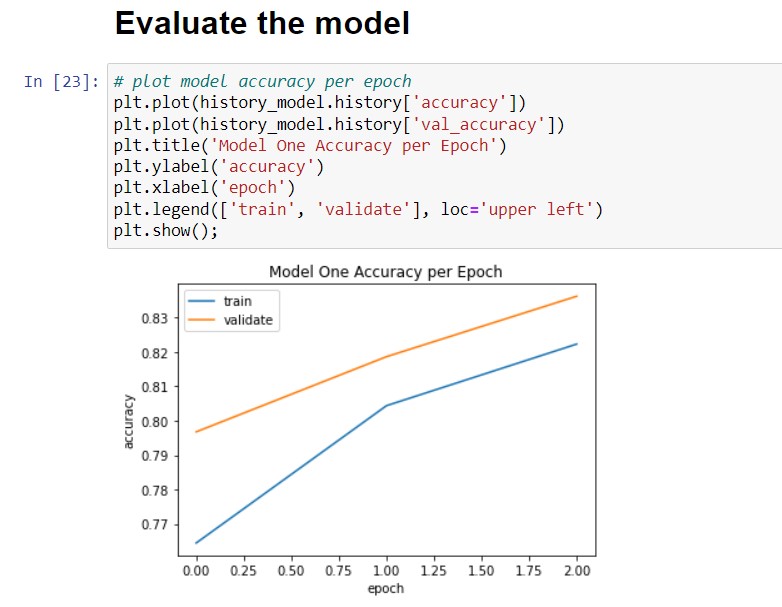
****

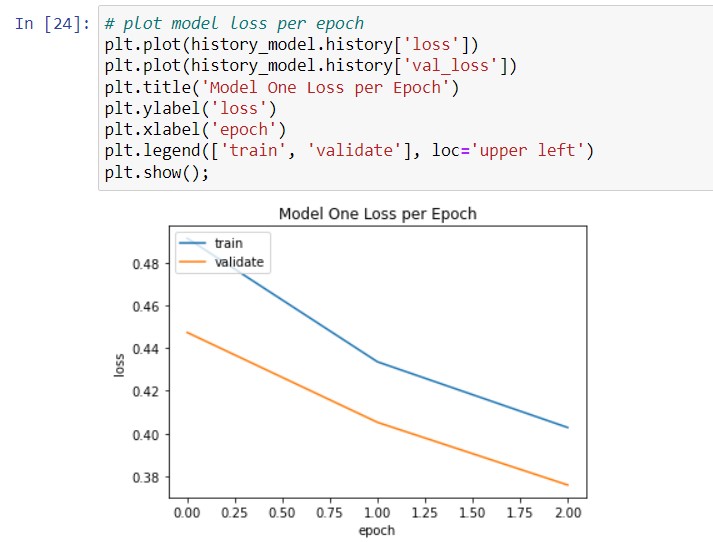
****

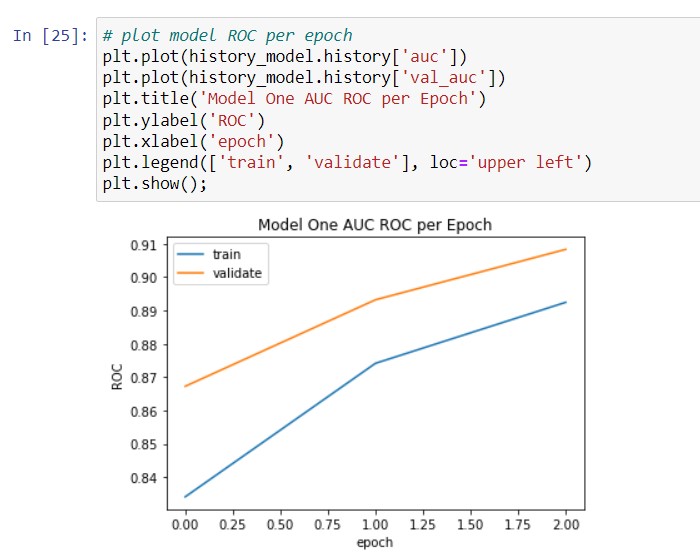
****

****

****

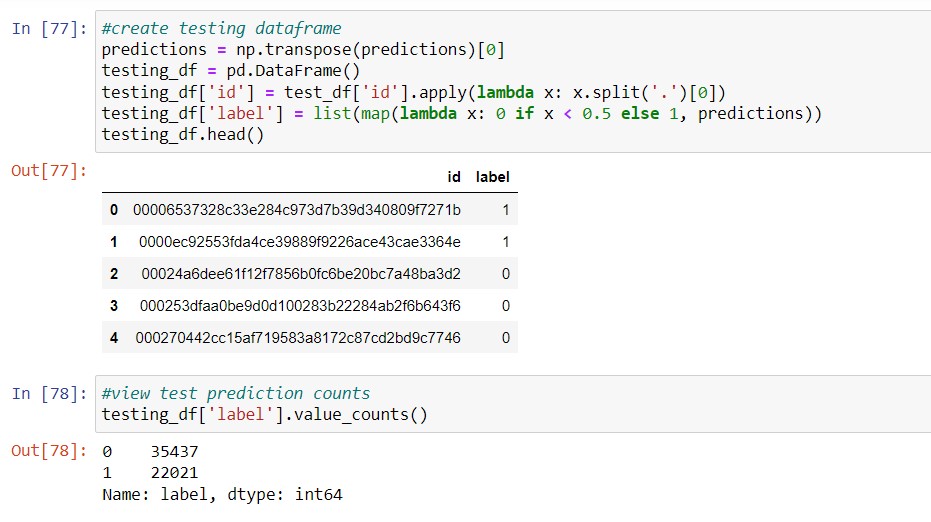
****

****

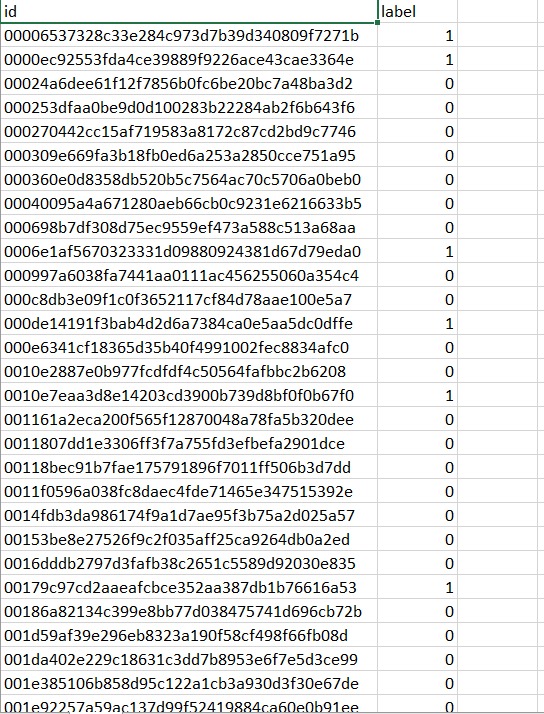
****

****

****

****

****

****

**10. RESULT**

From our test cancerous (1) cell images are 22021 and non-cancerous (0) cell images are 35437.

**11. CONCLUSION**

We can see from the above plots and diagrams for each model how well they performed with the training (and validation) sets. We see that model is a complex model with regards to the ROC metric. We see in the Model that the accuracy and loss do not steady, nor does the ROC in the model. This could pertain to the fact that we trained with very few epochs (3) and a simple CNN model with so many pictures may need more "time" to train to converge.

**12. FUTURE WORK**

As we know each body is different and exhibits different cancer symptoms every time. So each input data of all users are important in order to make the model more effective. In future we can use web application which based upon deep learning model it takes user input and recognise is blood cell is cancerous or not.

**13. REFERENCES**

* Edureka blogs and YouTube video
* https://www.kaggle.com/
* https://www.jhu.edu/