CHALLENGES FACED IN POTENTIAL USE OF BIG DATA ANALYTICS IN DETECTION OF BREAST CANCER

Dissertation Submitted to

The IIHMR University, Jaipur

for the partial fulfilment for the award of the degree of

Master of Business Administration (MBA)

In

Hospital and Health Management

By

Dr. KUNCHALA SAI SWAPNIKA

Under the Supervision and Guidance of

S.P. CHATTOPADHYAY

2021

CHALLENGES FACED IN POTENTIAL USE OF BIG DATA ANALYTICS IN DETECTION OF BREAST CANCER

Dissertation Submitted to

The IIHMR University, Jaipur

for the partial fulfilment for the award of the degree of

Master of Business Administration (MBA)

In

Hospital and Health Management

By

Dr. KUNCHALA SAI SWAPNIKA

Under the Supervision and Guidance of

S.P. CHATTOPADHYAY

2021

DECLARATION BY STUDENT

I hereby declare that this dissertation titled CHALLENGES FACED IN POTENTIAL USE OF BIG DATA ANALYTICS IN DETECTION OF BREAST CANCER is the bonafide record of my original research work. I declare that it has not been submitted to any other university or institution for the award of any degree or diploma. Information derived from the published or unpublished work of others has been duly acknowledged in the text.

Swapnika kunchala

(Signature)

Name of Student Dr. KUNCHALA SAI SWAPNIKA

Date 02.07.2021

CERTIFICATE

This is to certify that dissertation titled CHALLENGES FACED IN POTENTIAL USE OF BIG DATA ANALYTICS IN DETECTION OF BREAST CANCER is a record of the research work undertaken by KUNCHALA SAI SWAPNIKA in partial fulfilment of the requirements for the award of the degree of MBA (Hospital and Health Management) from the IIHMR University, Jaipur under my guidance and supervision and his/her approach to the research study has been sincere, scientific, and analytical.

Faculty Guide
IIHMR University, Jaipur

Date

School Dean
IIHMR University, Jaipur

Date

ABSTRACT

Artificial Intelligence is revolutionizing many industries. In the last decade many different sectors like online advertisements, autonomous vehicles, home systems, to a certain extent medical, etc have seen unprecedented changes. This paper explores some of the state-of-the-art AI techniques that are used to analyse images and how well they perform on mammography images in predicting if a patient has breast cancer or not and what are the major challenges that are faced in using these models in the health sector, especially in breast cancer detection and how the performance of these models can be improved. Different techniques that are studied are a simple convolution neural network, pretrained VGG19 and ResNet models that were conducted on 300 images (pretrained on 1.2 million generic images). As the data that was used to train the model was small pre-trained models VGG19, ResNet have shown around 92% accuracy and outperformed simple CNN model which has just 62% accuracy. With proper and sufficient data smart home assistant (Alexa), speech conversions, web search engines etc., are able to achieve 99% accuracy in their tasks. Similar results can be achieved in health sector for any task by gathering huge amounts of relevant data and training the models to attain the maximum benefit through technology.

ACKNOWLEDGEMENT

First and foremost, I would like to thank the IIHMR university for giving me this opportunity to study on a project and do the research. I would like to acknowledge the assistance of my mentor S.P Chattopadhyay who has guided me throughout the project and helped me not to miss the minute details.

I am indebted to my family and friends for making it possible for me to finish this thesis in such unprecedented times. Special thanks to my husband for being extremely supportive.

Finally, I humbly extend my gratitude to all those who have been directly or indirectly helpful in completing the thesis.

Table of Contents

1. INTRODUCTION

Over the last few decades, usefulness of data for better provision of service delivery has altered the way healthcare service has been perceived. Data is nothing but the collection of information which can be stored. The data may be used to solve many business problems and derive meaningful insights. Analytics can be used to filter massive amounts of data in seconds to find treatment options or solutions to different diseases. It is not only providing accurate solutions based on historical information but also provides customized solutions to unique concerns for patients.

Definitions

Artificial intelligence (AI) is a term used to describe a computer that performs jobs that are like performed by humans. Humans usually code or programme a computer for it to behave, reason, and learn. The code that informs the computer how to act, reason, and learn is known as an algorithm or model.AI is particularly good at finding patterns in vast amounts of data, deriving correlations between complicated aspects in the data, and discovering properties in data (such as photos) that the human brain cannot perceive.

Machine Learning (ML) is a sort of AI that can learn iteratively to make predictions or judgments rather than being explicitly programmed to execute a specific task. In the long term, the more data an ML model is exposed to, the better it will perform.

Deep Learning (DL) is a type of machine learning that uses artificial neural networks to learn from large amounts of data in the same way that the human brain does. In many instances, a well-designed and well-trained deep learning model may outperform human specialists at categorization tasks and making predictions.

Big Data is a word that refers to a large collection of data that continues to increase exponentially over time. Some of the examples of big data sources in healthcare sector are the hospital records, , medical examination results, patient medical records and internet of things devices. Big data analytics is a complex process of analysing massive amounts of data in order to uncover information such as hidden patterns and correlations so that better decisions can be made.

Biological Neural Networks (BNN):

A biological neural network is made up of several neurons that are connected to one another. Each neuron consists of a cell body, axon and the dendrites. Axons are cellular extensions that emerge from the cell body. Dendrites are narrow extensions of the cell body. Neurons receive information via dendrites and, like the axon, transmit them. Signals transfer from one neuron's axon to the dendrite of another at most synapses. The maintenance of voltage gradients in the membranes of all neurons makes them electrically active. The neuron produces an electrochemical pulse known as an action potential when the voltage changes by a large enough amount in a short period of time. As it reaches synaptic synapses, this potential travels quickly along the axon and activates them.

Artificial Neural Networks (ANN):

Artificial Neural Networks are inspired from the biological neural networks, they try to simulate functionalities that happen in our brain in a simple manner. ANN’s are a collection of layers that are connected, each layer has a set of nodes or units (each node like a neuron in our brain), one

layer relates to another with a set of weights. The input layer is on the top, the output layer is on the bottom, and the middle layers are concealed layers. ANN learns and tries to capture patterns in the data using sophisticated mathematical functions from the input data.

2. Review of Literature

Research studies have been conducted on how big data can detect patterns and turn high volumes of data into actionable knowledge in healthcare (Dash, S., Shakyawar, S.K., Sharma, M. *et al.* ,2019). A systematic mapping study was conducted to identify and analyse research on big data analytics and artificial intelligence in healthcare, in which 2421 articles between 2013 and February 2019 were evaluated (Mehta, Nishita & Pandit, Anil & Shukla, Sharvari,2019) The results of this study will help understand the needs in application of these technologies in healthcare by identifying the areas that require additional research.

Table 2.1 Publications relevant to ML methods used for cancer susceptibility prediction.

Publication	ML method	Cancer type	No of patients	Type of data	Accuracy	Validation method	Important features
Chen Y-C et al.	ANN	Lung cancer	440	Clinical, gene expression	83.5%	Cross validation	Sex, age, TNM_stage LCK and ERBB2 genes
Park K et al.	Graph-based SSL algorithm	Breast cancer	162,500	SEER	71%	5-fold cross validation	Tumour size, age at diagnosis, number of nodes
Chang S-W et al.	SVM	Oral cancer	31	Clinical, genomic	75%	Cross validation	Drink, invasion, p63 gene
Xu X et al.	SVM	Breast cancer	295	Genomic	97%	Leave-one-out cross	50-gene signature

						validation	
Gevaert O et al.	BN	Breast cancer	97	Clinical, microarray	AUC = 0.851	Hold-Out	Age, angioinvasion, grade MMP9, HRASLA and RAB27B genes
Rosado P et al.	SVM	Oral cancer	69	Clinical, molecular	98%	Cross validation	TNM stage, number of recurrences
Delen D et al.	DT	Breast cancer	200,000	SEER	93%	Cross validation	Age at diagnosis, tumour's size, number of nodes, histology
Kim J et al.	SSL Co-training algorithm	Breast cancer	162,500	SEER	76%	5-fold cross validation	Age at diagnosis, tumour's size, number of nodes, extension of tumour

Table 2.2 Publications relevant to ML methods used for cancer survival prediction.

Publication	Method	Cancer type	No of patients	Type of data	Accuracy	Validation method	Important features
Ayer T et al.	ANN	Breast cancer	62,219	Mammographic, demographic	AUC = 0.965	10-fold cross validation	Age, mammography findings
Waddell M et al.	SVM	Multiple myeloma	80	SNPs	71%	Leave-one-out cross validation	snp739514, snp521522, snp994532
Listgarten J et al.	SVM	Breast cancer	174	SNPs	69%	20-fold cross validation	snpCY11B2 (+) 4536 T/C snpCYP1B1 (+) 4328 C/G

3. Study Methodology

3.1 Research Questions

- What is the scope of big data analytics in determining cancer among patients?
- What are the challenges faced in adopting new models/tools of data analytics at each stage of cancer determination (diagnosis, risk prediction, treatment, modalities and prognosis)?
- what are the challenges faced in using the models to attain maximum accuracy of detection ?

3.2 Research Objectives

To find out and identify the obstacles that are causing hindrance in taking the complete advantage of using big data analytics in cancer treatment.

3.3 Research design- Descriptive analysis using Qualitative Methods

3.4 Research Methodology

The study is done through acquiring secondary data from published papers and related articles and building a sample model for better explanation. Hence the research is done following the descriptive qualitative methods.

Existing techniques that are used in healthcare: -

In the last three decades a variety of different ML techniques were used for disease prognosis and prediction. Major types of ML techniques that were used: Decision Trees (DT), Support Vector Machines (SVM), Bayesian Networks (BN's) and Artificial Neural Network (ANN's). Based on the data from two electronic databases PubMed, Scopus and focused on the cancer data which predicts susceptibility, recurrence, and survivability most of the existing research used simple ML models like DT, SVM, BN's. Genomic, clinical, histological, imaging, demographic, and epidemiological data were all contained in the input data.

The optimal platform for generating high throughput data sets is next-generation sequencing (NGS). It also necessitates the input of oncology experts with machine learning backgrounds in order to design algorithms for early-stage cancer detection through the identification of novel biomarkers and target sites, precise diagnosis through NGS sequencing, identification of selective target sites, and high-resolution medical imaging technology.

Fig 3.1 A single Neuron in BNN and ANN

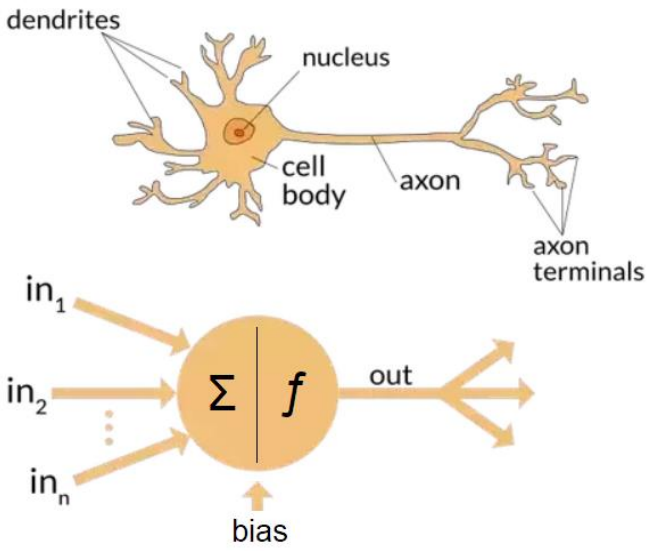
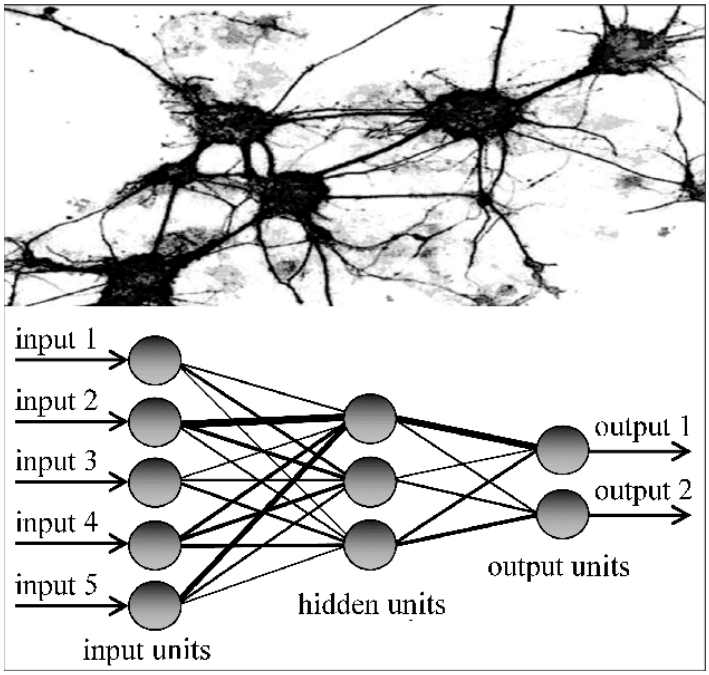


Fig 3.2 Comparison of Multi Layered ANN with BNN



Convolution Neural Networks (CNN's) in Health Care: -

Computer vision is a branch of computer science concerned with simulating human vision systems and allowing computers to process information from images and videos in the same manner that humans do. CNN is a special type of deep neural network that performs the task. Some of the popular applications that use CNN's are image classification, facial recognition, object detection along with its position in an image, image description, video motion analysis, image restoration and self -driving cars.

With a good amount of data CNN outperforms humans in many visual tasks, they are very good at predicting slightest of patterns in images, when trained with vast amounts of data 99%+ accuracy is not uncommon. Since CNN's perform exceedingly well with images data, in healthcare they can be used to analyse Mammography images, Histology images, X-Rays, MRI's, CT scans, etc.

A Study conducted using 3 popular CNN architecture: -

A study is performed on how well three different CNN models i.e., a simple CNN model, VGG - 19 model and ResNet model performed using tensorflow (a python package used for DL, AI) for the identification of tumour in the breast, if it is benign or malignant and results were interpreted below.

Simple CNN (Custom CNN model): -

Simple CNN's were the first popular neural network techniques that performed image classifications exceedingly well compared to other image classification techniques until this was introduced. It Should be trained using only given input data. With deep networks and large amounts of data we can achieve very good results.

Disadvantages: -

When we have a small amount of data identifying complex patterns with small architectures is not possible. When we have a small amount of data and using very deep networks can lead to memorization i.e., it performs very well on train data but does not perform well on seen data. With large amount of data and a very deep network it would take a lot of time to train the model. This model is Susceptible to vanishing gradient problem with a very deep network.

VGG 19

VGG 19 is one of the variants in VGG architectures (other is VGG 16). VGG network is trained on 1.2+ million images and each image is categorized into one of 1000 categories. It is First introduced in 2014 and has 19 layers in it. Pretrained ResNet model has pretrained weights in the initial layers, will takes advantages in identifying patterns that were already observed, has around 47 million parameters, 20 non trainable and 27 trainable parameters.

Disadvantages:

Very slow to train. Takes a huge amount of space to store the model. It is Susceptible to vanishing gradient problem with a very deep network.

ResNet

Deep Residual Network (ResNet) was arguably the most ground-breaking model in the deep learning community specifically in computer vision tasks. ResNet makes it possible to train up to hundreds and thousands of layers and achieve. It Has many popular variants ResNet18, ResNet50, Resnet101, ResNet500. Based on our input data we can choose the appropriate model. Pretrained ResNet model is also trained on 1.2+ million images. Even though the network is deeper the model size is significantly smaller than VGG models.

Disadvantages:

Very complex architecture. If sufficient data is available, and existing weights are not used, it would take a lot of time to train the model.

Mammography Images

The data is used from MIAS database. There are 330 mammography images stored in a png image format. Images are of low resolution 192 X 192 were used to build the model. As the images are less data augmentation techniques like rotation were used to build artificial data. Each image is classified as benign or malignant. Objective of the problem statement is given a patient mammography image identify the tumour is benign or malignant.

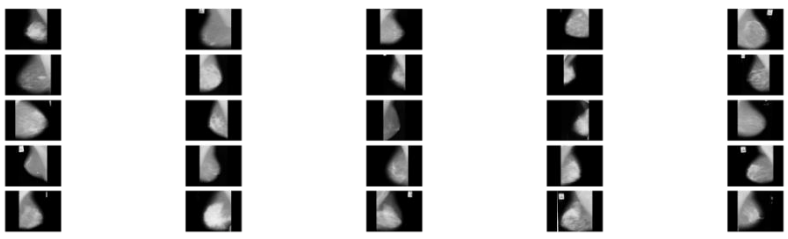


Fig 3.3 *Sample mammogram images data used for the study*

4 RESULTS

4.1 Simple CNN model

As we have very few images and we need to build a model from scratch, a deep neural network cannot be used, so performance of the model can be less compared to other powerful pretrained models. This model has an accuracy of 74 % on train data and 61.9 % on test data (validation or unseen data) and the error loss on test data is 0.89.

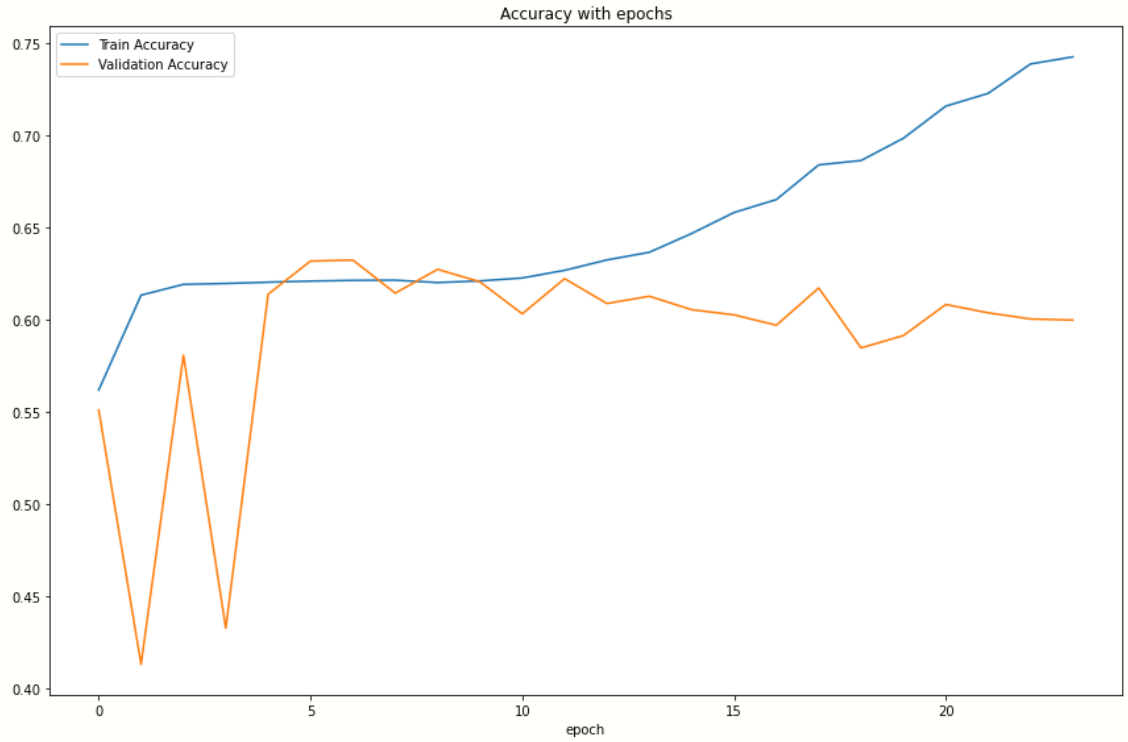


Fig 4.1.1

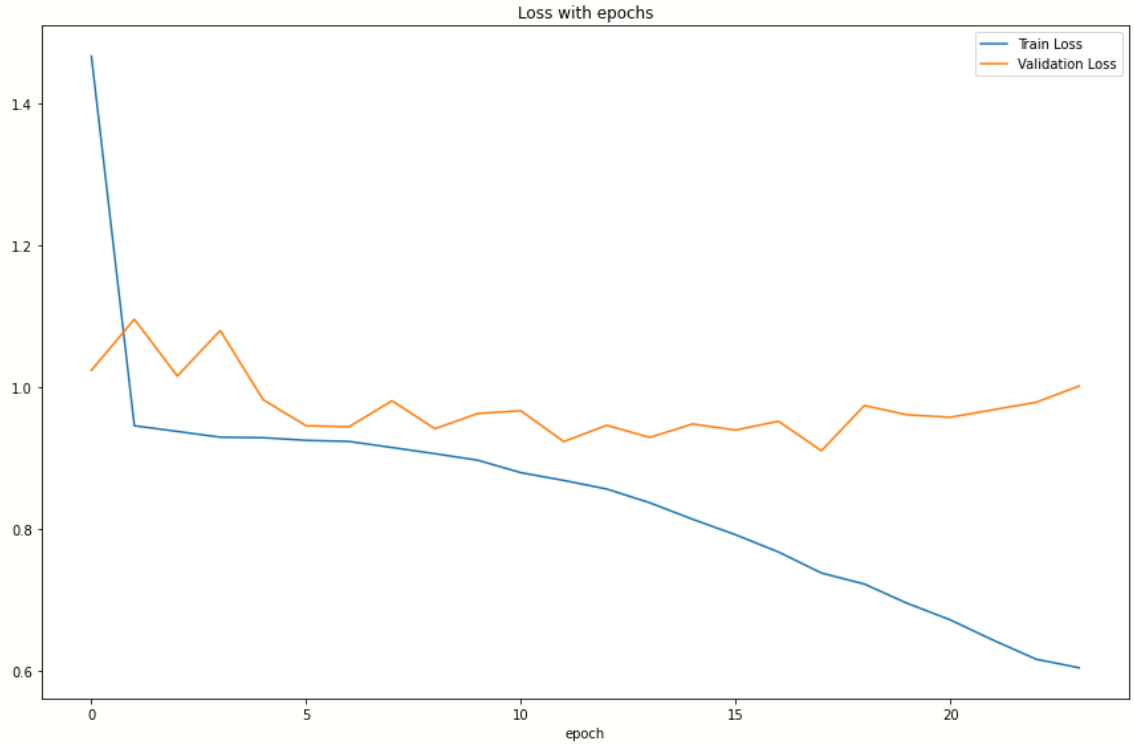


Fig 4.1.2

4.2 VGG 19 Model

Transfer learning is a very powerful and popular technique that can be used to build high performance and deep networks even when we have a small data set. When a CNN model which has VGG19 architecture and a few additional layers are added at the end an accuracy of 99 % was observed on train data and 93.6% was observed on test data. Also, it can be observed the error loss was 0.3 on test data which is significantly lower than that of plain CNN model.

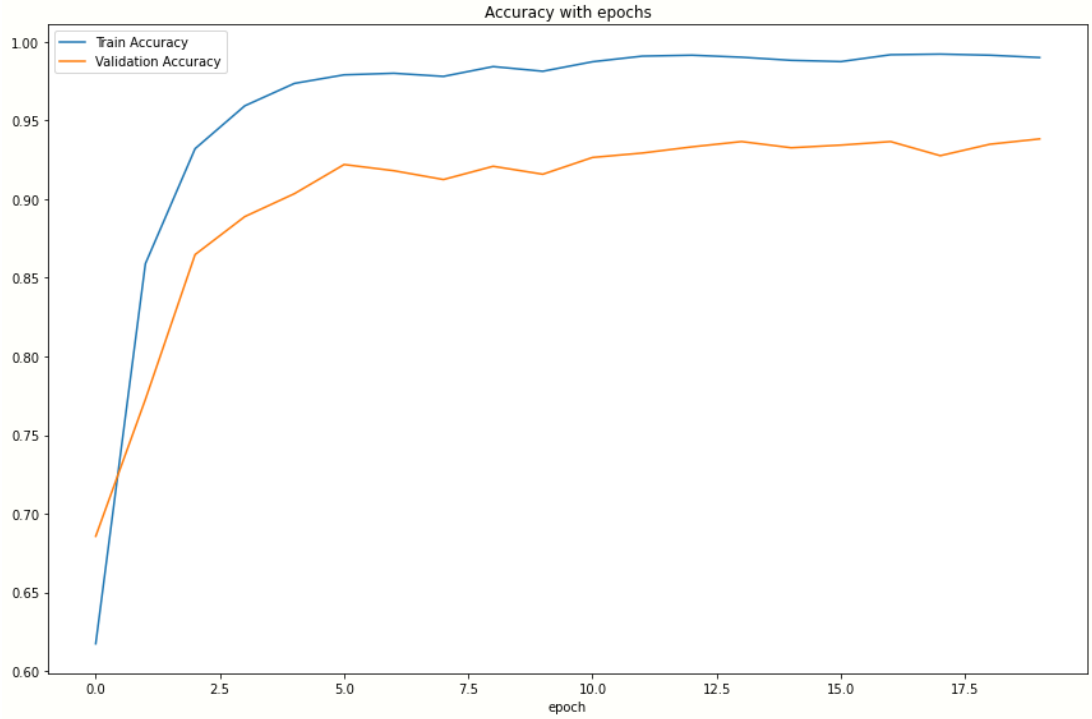


Fig 4.2.1

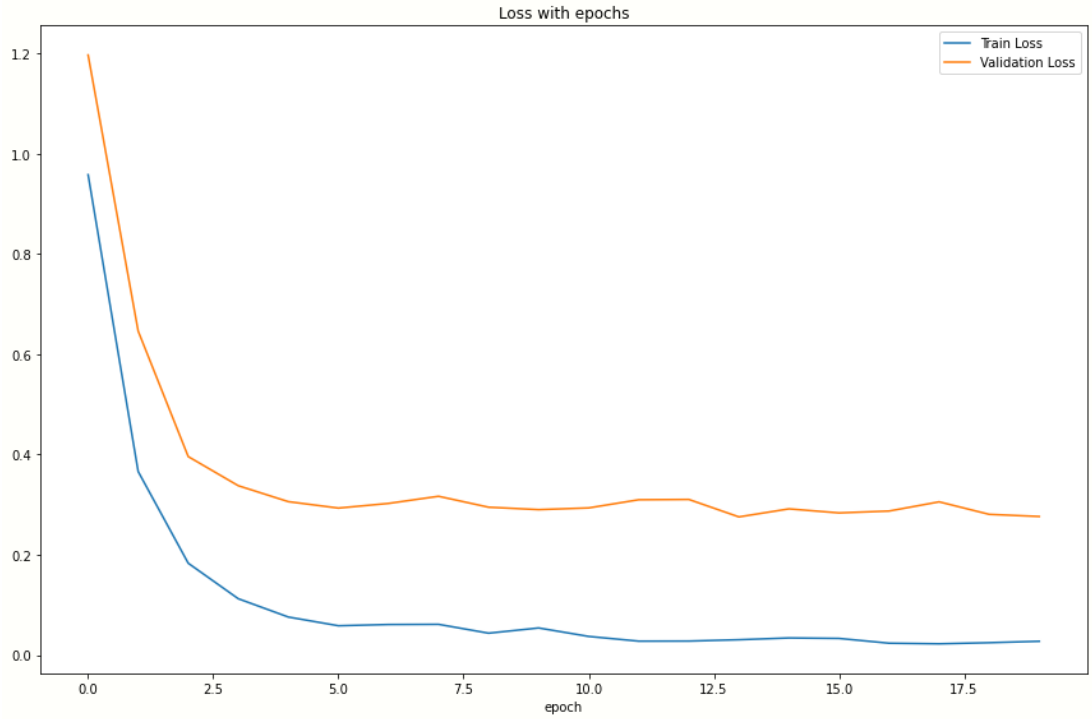


Fig 4.2.2

4.3 ResNet Model

When a ResNet50 model was used along with the transfer learning technique an accuracy of 99% was observed on train data and 91% was observed on test data. Error loss of 0.33 was observed on test data. A higher accuracy could have been achieved if more powerful ResNet101, ResNet500 models were used.

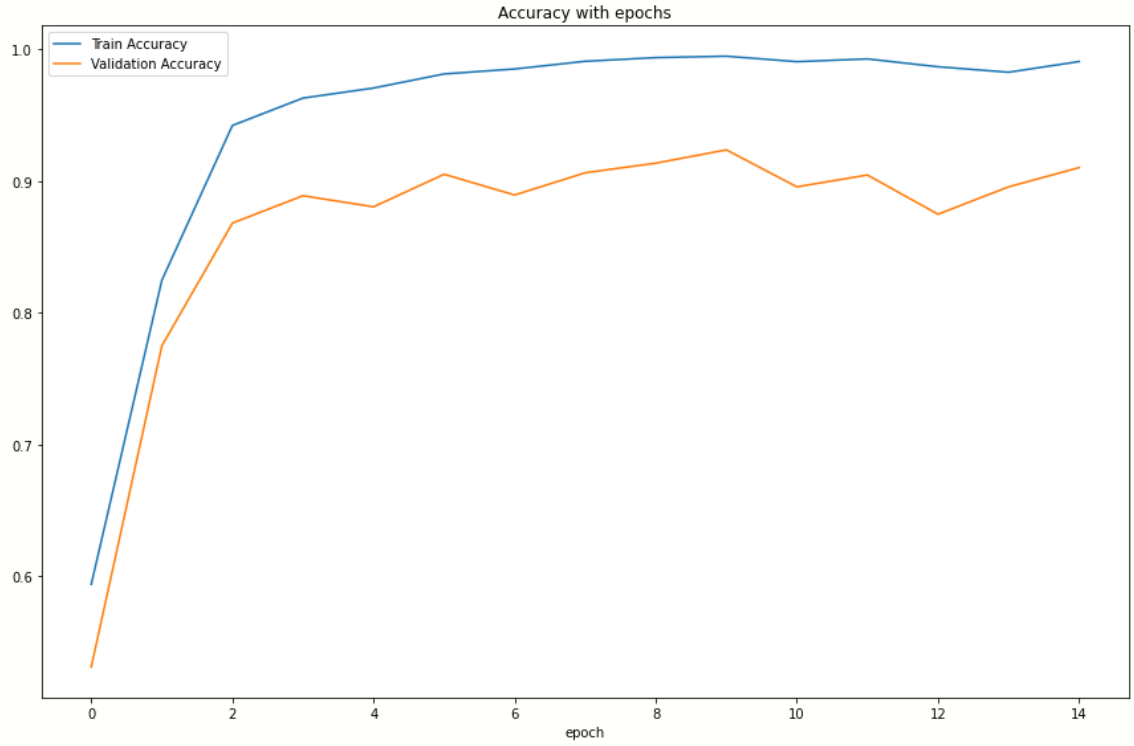


Fig 4.3.1

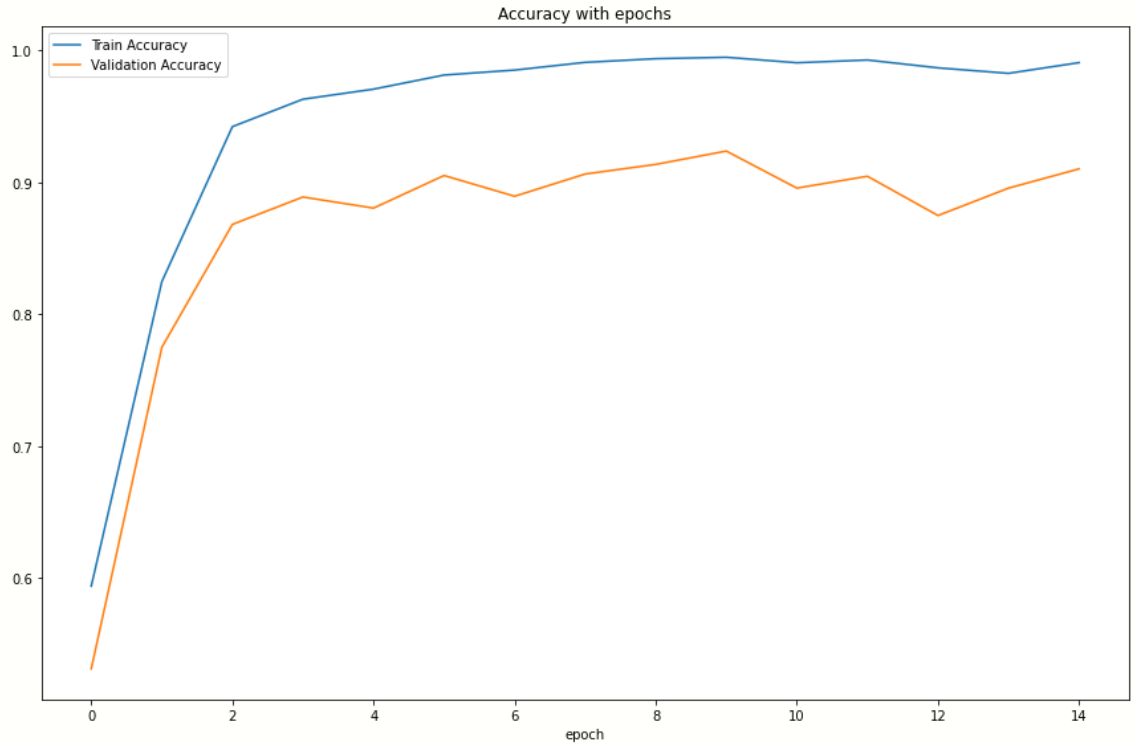


fig 4.3.2

5 FINDINGS

- Selection of appropriate Model

Most of the existing applications in healthcare and research that was published in healthcare have used data that has only images or raw data but not both. When all the relevant data points of an individual patient are collected, and this data is shared securely across the research community a super powerful model can be built.
In the study performed earlier, an accuracy of 93 % in VGG19 and 91 % in the ResNet 50 model was observed when pretrained models were used. These pretrained models was built using a random set of million plus images. Instead, if the model was trained using just the mammography images (one million plus) the accuracy could have been around 99%.

- Data collection:

Collection of specific data related to cancer is the most difficult part in the entire implementation and usage of these models. EHR data should be gathered and maintained for a longer period and then analysed. Data of existing cancer patients should be gathered along with those who are likely to be diagnosed with cancer in the family of the patient Out of which segregation of data specific to patients who are prone to acquire cancer is harder to drag from the whole set of undefined raw data of n number of patients. Data related to recurrence; prognosis of cancer should also be collected for analysis.

- Data analysis

As it is understood how the models work using a set of data to achieve results, sometimes it is not just enough for accurate prediction. In one of the models, images were used to read and identify the changes and predict cancer based on the difference and deviation from the normal images of a healthy patient. However, there are several instances where detecting cancer based on image analysis alone has not provided optimal results. In such cases, we need to join other data set which are essential along with images for better analysis.
For example, consider a patient has not shown any clinical symptoms and the images (mammograms, MRI, CT etc.,) were found to be normal, but the patient has genes which are specifically responsible for causing cancer. So, incorporating data related to hereditary details / reports over period of time along with image data for analysis provides better results than using only one data set.

- Manpower:

Though these models provide better results for cancer detection, however, for the process of data collection and analysis. Though the use of AI models reduces the manpower significantly, a shortage is usually observed in finding the people who are skilled in both clinical and technical domains.

6 CONCLUSION

Currently, there is a huge shortage of doctors in many countries across the world. With the advancements in AI better healthcare can be provided to the people and increase the lifespan of people. Now, even with little data AI techniques can be used to speed up the process of identification of a disease in a patient, which treatment would be better suited for a patient and achieve good results.

To use the best of AI in healthcare there are challenges that need to be used like data collection of patients and having a platform to share with the research community, people that are skilled in technical and in clinical domains, multiple data points that need to be collected for an individual person.

Despite the benefits of big data analytics in the healthcare sector, particularly in cancer care, worries regarding data collection and access, implicit and explicit bias, and challenges with trust in analytics technology among patients and providers have hampered its adoption in everyday healthcare. The more data a model encounters, the better it learns about the problem and is able to answer the query for which it was designed. As a result, large-scale data collection and analysis should be carried out in order to aid in the development of the most accurate models.

REFERENCES

1. https://www.cancer.gov/research/areas/diagnosis/artificial-intelligence
2. https://www.sciencedirect.com/science/article/pii/S2001037014000464
3. https://www.cancer.gov/research/areas/diagnosis/artificial-intelligence
4. http://peipa.essex.ac.uk/info/mias.html
5. https://healthitanalytics.com/news/4-emerging-strategies-to-advance-big-data-analytics-in-healthcare
6. https://docs.paperspace.com/machine-learning/wiki/convolutional-neural-network-cnn
7. https://arxiv.org/abs/1512.03385
8. https://resources.wolframcloud.com/NeuralNetRepository/resources/VGG-19-Trained-on-ImageNet-Competition-Data
9. https://github.com/swaposwap/bc_analysis_mammography_images
10. https://cancerci.biomedcentral.com/articles/10.1186/s12935-021-01981-1