Classification of Cancer using Epigenetic Markers in the Head and Neck

Author

MUHAMMAD OMAR ZEB

Regn # 00000204044

Supervisor

DR HASAN SAJID

DEPARTMENT OF ROBOTICS AND ARTIFICIAL INTELLIGENCE

SCHOOL OF MECHANICAL & MANUFACTURING ENGINEERING

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY

ISLAMABAD

June 2021

Classification of Cancer using Epigenetic Markers in the Head and Neck

Author

MUHAMMAD OMAR ZEB

Regn # 00000204044

A thesis submitted in partial fulfillment of the requirements for the degree of

MS Robotics and Intelligent Machines Engineering

Thesis Supervisor:

Dr. Hasan Sajid

Thesis Supervisor's Signature:

____

DEPARTMENT OF ROBOTICS AND ARTIFICIAL INTELLIGENCE
SCHOOL OF MECHANICAL & MANUFACTURING ENGINEERING
NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY,
ISLAMABAD
June, 2021

FORM TH-4

National University of Sciences and Technology

MASTER THESIS WORK

We hereby recommend that the dissertation prepared under our supervision by: **Mr. Muhammad Omar Zeb Regn#00000204044** Titled: **"Classification of Cancer using Epigenetic Markers in the Head and Neck"** be accepted in partial fulfillment of the requirements for the award of **MS Robotics & Intelligent Machine Engineering** degree. **(Grade_____)**

Examination Committee Members

1. Name: Dr. Yasar Ayaz Signature:________________

2. Name: Dr. M. Jawad Khan Signature:________________

Supervisor's name: Dr. Hasan Sajid Signature:________________

Date:____________________

__________________ ____________________

Head of Department Date

COUNTERSIGNED

Date:___________ ______________

Principal

Thesis Acceptance Certificate

It is certified that the final copy of MS Thesis written by *Muhammad Omar Zeb* (Registration No. 00000204044), of Department of Robotics and Intelligent Machine Engineering (SMME) has been vetted by undersigned, found complete in all respects as per NUST statutes / regulations, is free from plagiarism, errors and mistakes and is accepted as a partial fulfilment for award of MS Degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in this dissertation.

Signature: ___________________

Name of Supervisor: Dr. Hasan Sajid

Date: ___________________

Signature (HOD): ___________________

Date: ___________________

Signature (Principal): ___________________

Date: ___________________

Plagiarism Certificate (Turnitin Report)

This thesis has been checked for Plagiarism. Turnitin report endorsed by Supervisor is attached.

MUHAMMAD OMAR ZEB
Regn # 00000204044

Dr. Hasan Sajid
(Supervisor)

Declaration

I certify that this research work titled "*Classification of Cancer using Epigenetic Markers in the Head and Neck''* is my own work. The work has not been presented elsewhere for assessment. The material that has been used from other sources it has been properly acknowledged / referred.

Muhammad Omar Zeb
Reg# 00000204044

Copyright Statement

- Copyright in text of this thesis rests with the student author. Copies (by any process) either in full, or of extracts, may be made only in accordance with instructions given by the author and lodged in the Library of NUST School of Mechanical & Manufacturing Engineering (SMME). Details may be obtained by the Librarian. This page must form part of any such copies made. Further copies (by any process) may not be made without the permission (in writing) of the author.
- The ownership of any intellectual property rights which may be described in this thesis is vested in NUST School of Mechanical & Manufacturing Engineering, subject to any prior agreement to the contrary, and may not be made available for use by third parties without the written permission of the SMME, which will prescribe the terms and conditions of any such agreement.
- Further information on the conditions under which disclosures and exploitation may take place is available from the Library of NUST School of Mechanical & Manufacturing Engineering, Islamabad.

Acknowledgements

I am thankful to Allah Almighty for bestowing me with the energy and will that enabled me to successfully complete this project.

I am thankful to my supervisor Dr Hasan Sajid and the School of Mechanical and Manufacturing Engineering who supported me in every possible way throughout the course of my thesis.

Dedicated to my parents and friends without whom support I would not have achieved such an accomplishment.

Abstract

Using the DNA methylation data present in The Cancer Genome Atlas, we propose a new data preprocessing method where we use the caner driver genes to extract the relevant features from the data. After the preprocessing step we performed a feature extraction method where we selected top 50 features from each of the four sites of the human body. This method of feature extraction method yielded a comparable F-score against other studies while also reducing the overall space complexity of the problem

Keywords*: DNA Methylation, Driver Genes, mRMR, TCGA, Cancer, Feature Extraction, Machine Learning, Neural Networks*

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Description
TSG	Tumor Suppressor Gene
DNA	Deoxyribonucleic acid
TCGA	The Cancer Genome Atlas
CNN	Convolutional Neural Network
GEO	Gene Expression Omnibus
ROC	Receiver Operative Characteristic
mRMR	Maximum Relevance Minimum Redundancy
ANN	Artificial Neural Network

CHAPTER 1 INTRODUCTION

1.1 Background:

Cancer is the second leading cause of death in the world and accounted for 9.6 million deaths in the year 2018, worldwide, first being deaths caused by heart diseases. The most fatal cancers in descending order are lung, stomach, and breast. The field of cancer diagnostics has seen some major changes in the past few years owing to the rapid evolution in computing. Advanced and continually improving techniques in data science and statistics, coupled with ever-growing computing capabilities has enabled researchers to now use large amounts of genetic data for diagnosis, prevention, and treatment of various types of cancers.

Studies in bioinformatics and genomics have established that Deoxyribonucleic Acid (DNA) methylation data is significantly more reliable at predicting cancer than the gene expression data. DNA methylation is the process in which a methyl group (CH_3) gets attached to the gene sequence. DNA methylation alters the transcription of the gene without changing the sequence of the gene, thus this process is categorized into the study of epigenetics. Various studies have shown that a very specific set of genes is responsible for the transformation of healthy cells into cancerous ones, among them Tumor Suppressor Genes (TSGs) and Oncogenes are the most noteworthy. While other genes might passively impact the process, they do not actively partake in the process of said transformation and are passenger genes. TSG, as the name implies, is responsible for the inhibition of the growth of a cell but when these genes become inactivated the chances of getting cancer get higher. Oncogenes, on the other hand, are responsible for the growth of the cell, their activation results in the abnormal growth of the cell, causing the cell to become cancerous. Study shows that the promoter regions of the TSG are generally hyper methylated causing it to be suppressed and Oncogenes to be hypo methylated causing them to be abnormally active in cancerous cells.

1.2 Aims and Objectives:

The aim of the is work is to design and develop a classification model using deep learning which can accurately classify benign and malignant tumors with high accuracy and can also classify the location to which the tumor is associated.

1.3 Research Methodology:

This work is divided in to following steps:

i) Data extraction from The Cancer Genome Atlas (TCGA) online repository.
ii) Feature extraction.
iii) Developing artificial neural network for binary and multiclass classification.
iv) Using neural network and feature extraction method to reduce the number of features

CHAPTER 2 LITERATURE REVIEW

2.1 Database:

The DNA methylation data used in this study was extracted from The Cancer Genome Atlas (TCGA) database. TCGA project was started in 2005 by National Cancer Institute funded by the United States Government. It started as a project to characterization of three cancer types in the human body. After 11 years the number of cancer types in the database has increased to 33 which also include 10 rare cancers. The database is now split into two parts:

1) Genome Characterization: it deals with genome sequencing.
2) Genome Data Analysis: it deals with bioinformatics analysis.

The data present in the database includes gene expression, copy number variation, single nucleotide polymorphisms, DNA methylation, microRNA profiling and exon sequencing.

In this study DNA methylation data was used as it is more accurate in predicting the type of cancer as compared to gene expression data.

The DNA methylation data used in this study was extracted using Illumina Infinium HumanMethylation450 Bead Chip. The Bead Chip is a cost-effective and high throughput assay capable of extracting the methylation data from about 450,000 CpG sites, providing a high resolution of epigenetic changes. A CpG site refers to location in the DNA molecule where cytosine (C) nucleotide is followed by guanine (G) nucleotide.

2.2 Previous Work:

As the database used in this study is open-source and epigenetic markers are more reliable in identifying the cancers, the database has been used in studies to classify between benign and malignant tumors. Four research works were studied before starting the project.

The first study was targeted against breast cancer classification as it is the major cause of death among women. This study combined both DNA methylation and gene expression data to improve classification model. The study proposes that by combining gene expression and methylation data the results would be able to see the differences on transcript and epigenetic level. The proposed model of random forest was able to classify the cancer with an error of 0.1 and 0.5 depending on the subtype of cancer. The proposed model in this study was able to identify some unique data points in the methylation data which were previously not considered. [2]

Due to the use of random forest as a feature selection method the above study finds it very difficult to select the features for certain types of Brest cancer notably HER2. This might be since the data provided to the feature extraction model was highly imbalanced and there was an overlap between the selected features which might suggest that the feature selected were not adequate to accurately classify the cancer sub-types in the Breast.

The second study was focused on breast, kidney and thyroid cancers. This study used a random forest classifier. The study uses an iterative method of using random features as input to the random forest and kept on iterating until the threshold of accuracy was crossed. The threshold of accuracy was set to 98% on the test set. Once the threshold was crossed the model was stopped and outputted the features on which the model was trained. Those features were then mapped to the genes to give an insight on which genes are relevant to the cancer. [3]

This study achieves high accuracy using the proposed method but due to use of iterative method it is not only slow to converge due to high number of features present in the data but is also very compute intensive as a lot of decision trees must implemented to look for the right feature set. This method of classification was implemented on an Apache Spark Cluster with 66 nodes, and it took around 1 hour to train the model and 20 minutes to run the inference.

The third study used cancers related to 33 different sites present in the human body. In the study a novel method was proposed where the methylation dataset was converted in to 2-dimensional matrix of size 220x1663, where each value of the matrix defined the methylation level of a CpG site. The matrix was then fed into a convolution neural network (CNN) having both height and width wise filters. The accuracy achieved on the test set, by using this method, was 92.87. This study was used to identify the region to which a specific cancer was associated to. [4]

This method produces comparable results to the other studies mentioned but by converting the data to a 2D image most of the useful information is lost as the 2D image created using this image resembles to that of salt and pepper noise. Using this method of classification, we cannot determine which genes were contributing to the cancer causation as no feature selection method was applied beforehand.

The final study used the Lungs data from the TCGA database and Gene Expression Omnibus (GEO) database, and the data was based on the Illumina HumanMethylation27 assay rather than the HumanMethylation450 data as used in the other works. This work proposes an ensemble model comprising of multi-category receiver operating characteristic (Multi-ROC), random forest and maximum relevance and minimum redundancy (mRMR). The Multi-ROC and mRMR were used as a feature selection method while random forest was used as a classifier model. Using the ensemble model the study reported accuracy of 84.60% on an independent test set.[5]

Due to the use of two different feature selection method and choosing only those features which are common in both feature selection method the space complexity of the problem is reduced using this method, but it increases the time complexity of the problem as two different algorithms must be used before the training step to extract the relevant features. Also, by looking at the accuracy of the overall algorithm the algorithm is performing not so well as compared to other studies that might be since the number of features that are overlapped in the feature extraction method are small which is affecting the overall accuracy.

All the studies described in this section either used all the features present in the data or used some sort of data manipulation to reduce the number of features. This step is fine but has one drawback that not all genes contribute towards the cancer causation. The genes that promote cancer are called cancer driver genes while others are called cancer passenger genes.

2.3 Cancer Driver Genes:

Comprehensive studies in the genomic sequencing have developed a landscape of the human cancers which consists of mountains and hills. The mountains define the genes which are altered in high percentage of cancers and hills define the genes which are altered in low per percentage of cancers.

Studies have found that mountains are made up of approximately 140 genes. As these genes promote the causation of cancers, they are called driver genes. A typical cancer in the human body is caused by the mutations of about two to eight of these driver genes. The driver genes that have been identified are responsible for three of the core cells processes:

1) Cell fate: this defines the identity of the daughter cell.
2) Cell survival: this defines the regulated or unregulated death of a cell.
3) Genome maintenance: this defines the DNA repair and cell division cycle.

The driver genes are split into two parts:

1) Tumor Suppressor Genes
2) Oncogenes

Tumor Suppressor genes are responsible for the inhibition of the growth of the cells. When these genes get suppressed, there is a high chance that the cell will have an abnormal cell division and as the driver genes are responsible for the cell fate, this will affect the daughter cells as well and a tumor will form.

Oncogenes are responsible for the cell division. When these genes get hyper expressed, they can also disturb the cell division and as is the case with the tumor suppressor they will affect the daughter cells as well and a tumor will form.

The genes that define the hills and valleys in the landscape are called the passengers genes. The mutations in these genes do not contribute to the causation of the cancer in the human body.

CHAPTER 3 METHODOLOGY

3.1 Data:

TCGA database was used to acquire the DNA methylation data of four regions associated with the human body. The four regions are:

1) Lungs
2) Breast
3) Stomach
4) Head and Neck

A total of 2702 samples were downloaded from the database. The numbers of samples per region are shown in table [1].

Region	No. of samples
Head and Neck	479
Lungs	827
Breast	795
Stomach	405
No Cancer	196

Table 1 Number of samples per region

Along with the data of cancerous cells, the database also contains some samples which contain methylation data of non-cancerous part in the same region of the same patient.

Each sample of data is a .csv file which corresponds to a single patient with an associated .txt file containing the information about the tumor type i.e., benign or malignant. The .csv file has 485,578 rows and 11 columns. Each row corresponds to a single CpG site while the columns contain different information about the site.

In this study not all columns were used as most of the columns contained information that was same for all the data irrespective of the type of cancer or the region.

The columns that were used in this study are:

1) Start: This defines the starting point of the CpG site
2) Beta Value: This defines the methylation level at the CpG site.
3) Gene Symbol: This defines the associated gene of the CpG site.

As all the data was acquired from the Illumina Infinium HumanMethylation450 Bead Chip, the start and the gene symbol columns contained similar data across all the samples only the beta values differed in all the samples.

The beta value is defined as the ratio of methylated intensity and the overall intensity and is calculated using the equation [1].

$$\beta = \frac{max\ (y_{i,meth}, 0)}{max(y_{i,meth}, 0)\ + max(y_{i,unmeth}, 0)\ +\ \alpha} \quad [1]$$

Where y_(i,meth) and y_(i,meth) are the intensities measured by the ith methylated and un-methylated probes, respectively. α is used to regularize the Beta value. The default value of α is 100. Beta Value is always between 0 and 1. Beta value 0 means that the CpG site is hypo methylated and beta value 1 means that the CpG site is hyper methylated.

3.2 Data Preprocessing:

As already identified in [4], the passenger mutations have no direct effect on the growth advantage of the cell and are not cancer causing. On the other hand, driver gene mutations play a significant role on the growth advantage of the cell which can lead to cancer. There are about 125 driver genes comprising of 54 Oncogenes and 71 Tumor Suppressor Genes present in the human genome [4].

Based upon the above-mentioned studies, we propose a method to extract the beta values from the patient's data which only utilizes those genes that are responsible for the driver gene mutations. Our technique specifically, searches for driver Gene Symbols in the patient's data (hence the use of Gene Symbol column) and copies the starting point associated with the driver gene of the CpG site and the associated Beta Value for that starting point. Using this extraction method, the number of features is reduced from 485,578 to just 296, which is a tremendous improvement in terms of space and time complexity. This method also highlights the fact that there are only 38 driver genes present in the data, which are majorly responsible for causing cancer.

Once the driver genes were identified and their associated CpG site and beta values were extracted, the gene symbol was discarded.

The NA values in the data are converted to zeros and all the data is combined to form a single .csv file with rows identifying the patient and the columns identifying the Beta Value of the CpG site taken as a feature. As all the sample data was generated using same method, the starting points were identical across all the samples.

The data was then divided into training and testing sets with training data comprising of 90% of the data and testing set comprising of 10% of the whole data using random sampling.

The data distribution was done twice, once for multi-label classification problem with five labels one for each region and fourth for no cancer and second for binary classification problem with 1 defining cancer and 0 defining no cancer label.

The data preprocessing step was done in MATLAB 2018b.

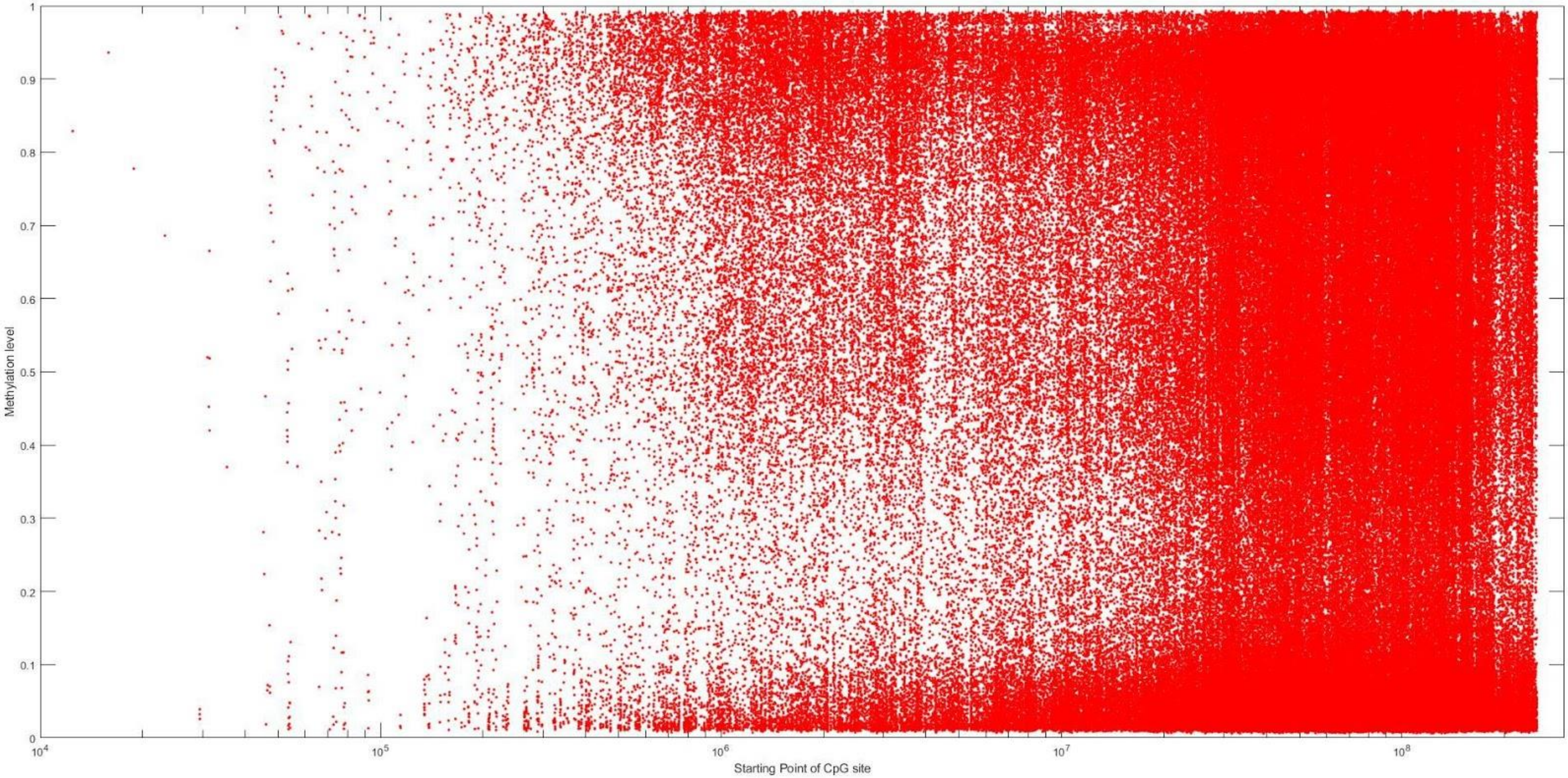

Figure 1 Data plot containing all the data points of a single patient.

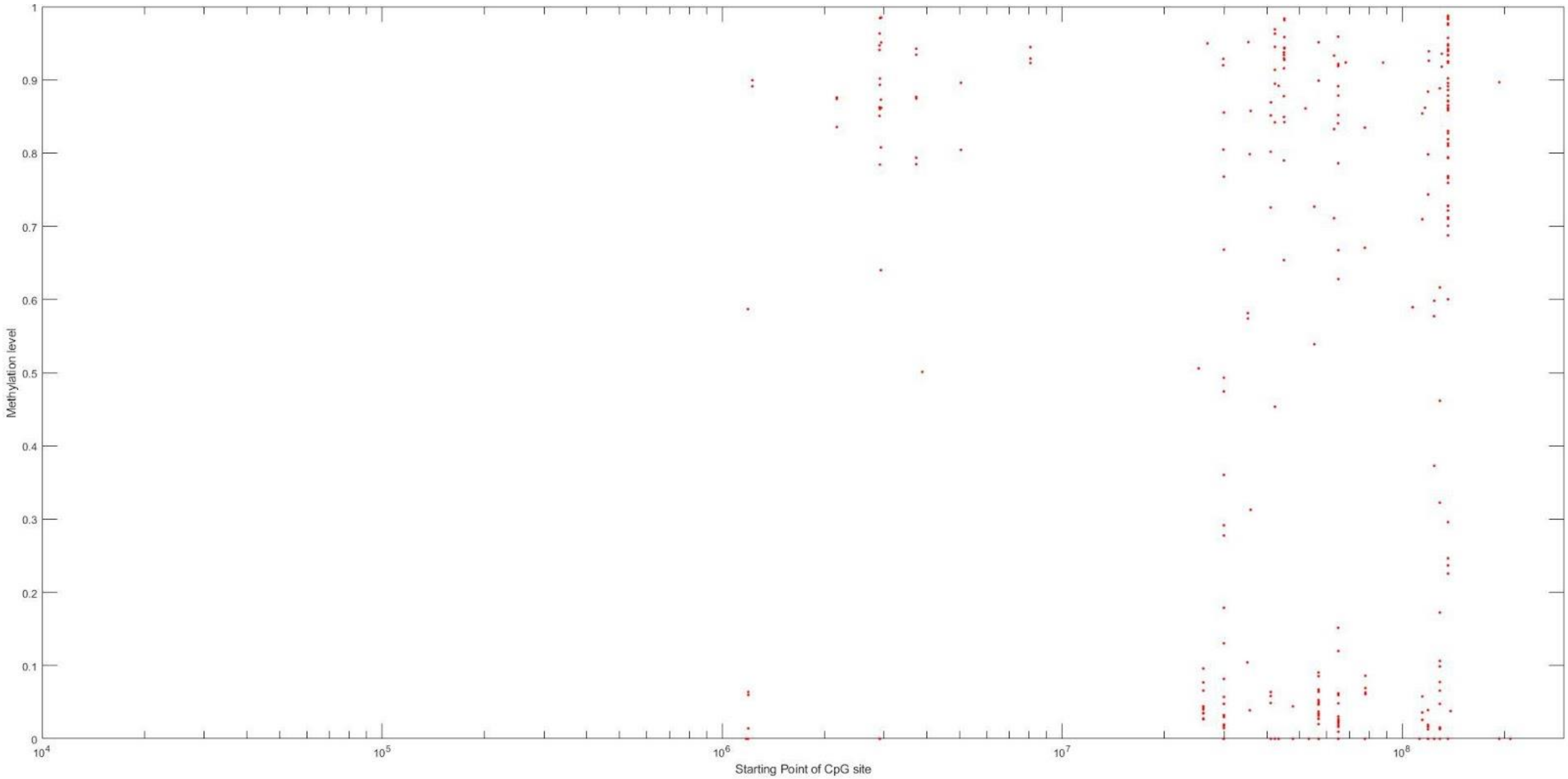

Figure 2 Data plot containing only driver genes of a single patient.

3.3 Classification Model:

In this study, a simple neural network was developed for the classification using keras library in python 3.7. The neural network consists of 4 layers with relu activation function equation [2] in the first three layers and a softmax activation function equation [3] for the multiclass problem and sigmoid activation function equation [4] for binary class problem, in the last layer. The layers were initialized using Xavier Initializer with a seed value of 10.

The neural network was optimized using RMSprop equation [5 – 8] with the beta value of 0.9 and learning rate of 0.00001. For the loss function categorical cross entropy equation [9] was used and the model was trained using 1500 epochs.

To measure the performance of the model F-Score equation [10] and confusion matrix were used.

$$R(x) = \begin{cases} x, & x > 0 \\ 0, & x \leq 0 \end{cases} \quad [2]$$

$$f(x) = \frac{e^{x_i}}{\sum_{j=1}^{K} e^{x_j}} \quad [3]$$

$$f(x) = \frac{1}{1 + e^{-x}} \quad [4]$$

$$v_{dw} = \beta \cdot v_{dw} + (1 - \beta) \cdot dw^2 \quad [5]$$

$$v_{db} = \beta \cdot v_{dw} + (1 - \beta) \cdot db^2 \quad [6]$$

$$W = W - \alpha \cdot \frac{dw}{\sqrt{v_{dw}} + \epsilon} \quad [7]$$

$$b = b - \alpha \cdot \frac{db}{\sqrt{v_{db}} + \epsilon} \quad [8]$$

$$Loss = -\log\left(\frac{e^{s_p}}{\sum_{j}^{C} e^{s_j}}\right) \quad [9]$$

$$F\ Score = \frac{2 * Precision * Recall}{Precision + Recall} \quad [10]$$

The F-Score for the multi-label problem was calculated by taking the indices of the maximum value along the row axis using the np.argmax() function of numpy library.

The model was trained and tested using the computer with specification given in table [2].

Parameter	Value
Architecture	X86
CPU	Intel® Core™ m3-7Y30 @ 2.60 GHz
No. of cores	2
No. of threads	4
RAM	8 GB
OS	Windows 10

Table 2 Computer used for experiments.

The complete summary of the model used in training and testing of the data is given in table [3] and table [4].

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 148)	14060
dense_1 (Dense)	(None, 74)	11026
dense_2 (Dense)	(None, 37)	2775
dense_3 (Dense)	(None, 5)	190

Table 3 Model summary for multi-label problem

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 148)	14060
dense_1 (Dense)	(None, 74)	11026
dense_2 (Dense)	(None, 37)	2775
dense_3 (Dense)	(None, 2)	76

Table 4 Model summary for binary class problem

3.4 Feature Extraction:

Using the above-mentioned model and data processing techniques we were getting comparable results with other studies done on the same data. Once we finalized our model, we switched towards extracting the features which were most prominent in a given cancer site. For feature extraction method we used maximum relevance minimum redundancy (mRMR) feature selection method.

3.4.1 Maximum Relevance Minimum redundancy:

Feature selection methods can be categorized into three categories:

1) Embedded Method
2) Wrapper Method
3) Feature Based Method

Embedded method of feature selection uses an iterative penalization function along with the machine learning algorithm to find the feature set that is most suitable for that specific task.

Wrapper method uses a subset of features from the total set of features and then decides based on the scores of the test set, to keep which set of features.

Feature based method is independent of the model and can be described as a pre-processing set where different types of statistical analysis are done on the feature set to reduce the number of features.

The mRMR feature extraction method falls into the category of feature-based method as it does not need any model to find the best possible feature set.

The mRMR feature extraction method works by finding those features which have the highest correlation with the target label while having the lowest correlation with each other. The inputs of the algorithm are then ranked based on the said criteria.

The correlation between the target label and feature and between features are calculated using the mutual information equation [11].

$$I(x,y) = \sum_{i,j} p(x_{i,} y_i) log \frac{p(x_i, y_i)}{p(x_i)p(y_i)} \text{ [11]}$$

In the above equation x and y are the two random variables whose correlation we need to calculate. $p(x_i, y_i)$ is the joint probability distribution of the two variables and $p(x_i)p(y_i)$ are the marginal probabilities of the variables.

Using the equation [11] we can find the relevance between the target class and the variable by using equation [12].

$$V(S) = \frac{1}{|S|} \sum_{i \epsilon S} I(y, i) \text{ [12]}$$

In the above equation S is the number of features that we need to select, y is the target class and i is the feature. The goal of this equation is to find those features which have the maximum V(S) value.

Using the equation [11] we can find the redundancy between the two variables by using equation [13].

$$W(S) = \frac{1}{|S|^2} \sum_{i,j \epsilon S} I(i, j) \text{ [13]}$$

In the above equation S is the number of features that we need to select, i and j are the features. The goal of this equation is to find those features which have the minimum W(S) value.

In order to find the features which, have maximum correlation to the target class while having the minimum correlation between the features we can combine equation [12] and equation [13] to get the final output which is given by equation [14- 15].

$$mRMR(S) = \max\,(V(S) - W(S)\ [14]$$

$$mRMR(S) = \max\left(\sum_{i\epsilon S} I(y,i) - \frac{1}{|S|}\sum_{i,j\epsilon S} I(i,j)\right)[15]$$

3.4.2 Feature selection:

As the mRMR feature selection requires the number of features to be outputted, we ran the mRMR feature selection method in tandem with our machine learning model to find the approximate number of features which when given to the neural network gives high F1 score on the testing set.

In the first iteration of the feature selection method, we selected the number of features to outputted by the mRMR method to be 100. Once the top 100 hundred features were selected having high relevance to the target class and minimum correlation with each other's. The method was run independently on each site with the target class given as cancer and non-cancer. Once all the features were selected, the features of all the sites were combined and the neural network was trained and tested 5 times and the average F-score was calculated. Once this was done the number of features were reduced by a factor of 25, the above method was repeated and the average F1 scores were calculated. This was repeated 4 times till reached the total number of 25 features. Figure [3] plots the average F-scores against the number of features.

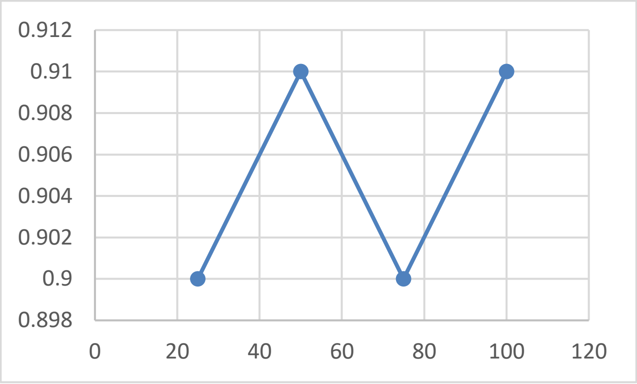

Figure 3 F-score vs number of features.

From figure 3, we can see that the result for 50 features and 100 features is the same so we can select either one of them. As a smaller number of features means less computational power is needed so we selected the 50 features.

To further validate the number of features selected we then repeated the above-mentioned method once again but this time we started from 55 and reduced the features with a decrement of 5 until we reached 40 features. The average F Score for this method is given in figure [4].

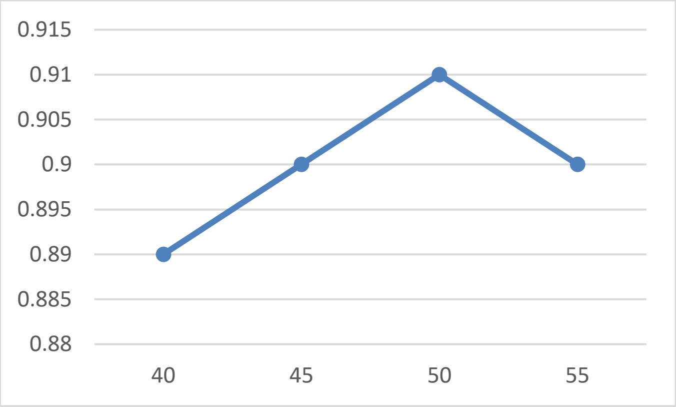

Figure 4 F-score vs number of features.

From the figure 4 we can see that the best scores are achieved when we select 50 features. Using this knowledge in hand we selected top 50 features from each site.

This mRMR feature extraction method was implemented using pymRMR Feature Selection library in Python.

CHAPTER 4 RESULTS AND DISCUSSIONS

4.1 Results:

The artificial neural network model was first trained and tested on the data that was extracted using the cancer driver genes using the method proposed in section 3.2. Using the preprocessing method, the data points comprised of 296 features and 38 genes per patient.

The artificial neural network model was trained and tested 5 times and average F score was calculated. The model was trained for 1500 epochs in each iteration. This was done for both multi-label and binary classifications. The F Score for multi-label classification was reported as 0.92 and F-Score for binary classification was reported 0.98 on the test set. Figure [5] shows the training

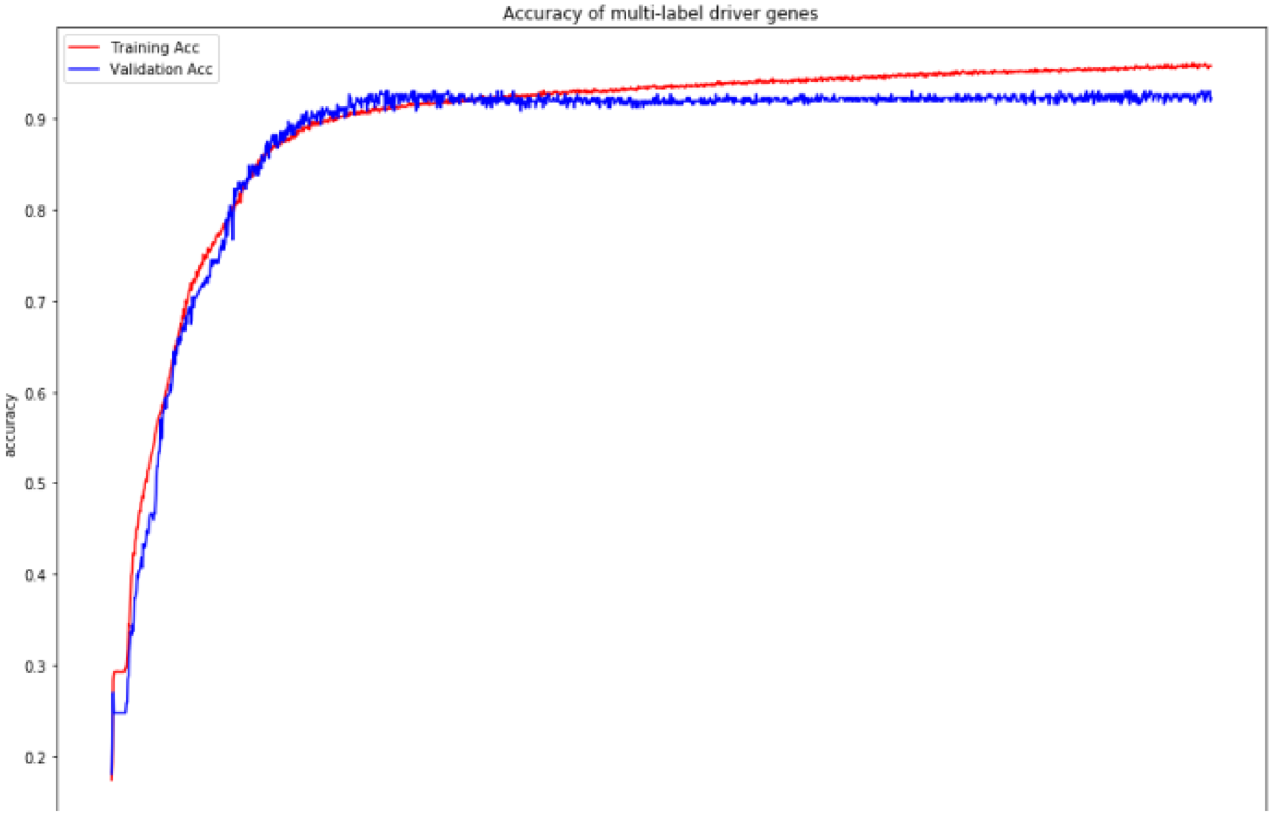

Figure 5 Training and Validation Accuracy for Multi-class problem using Cancer Driver Genes.

and validation accuracy for the multi-class problem and Figure [6] shows the training and validation accuracy for the binary-class problem for each epoch.

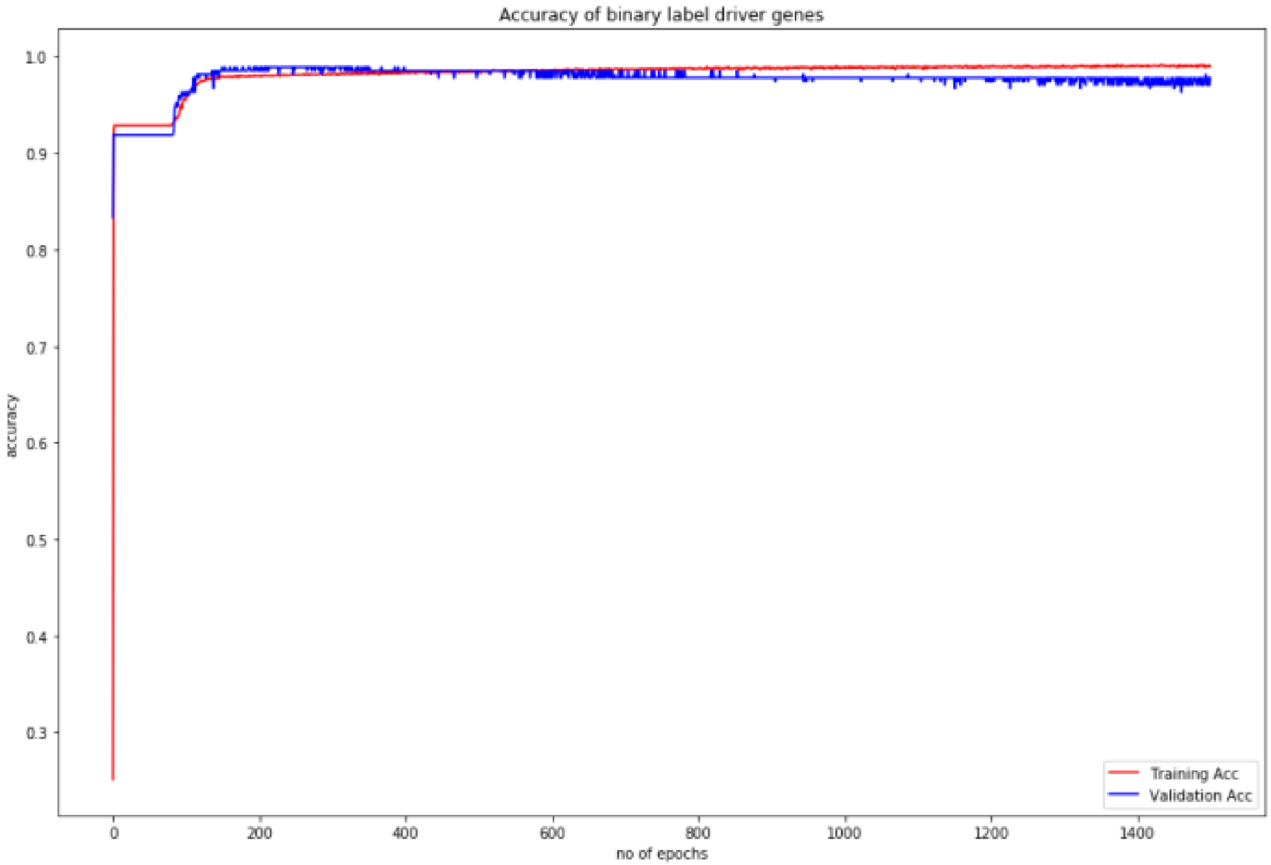

Figure 6 Training and Validation Accuracy for Binary-class problem using Cancer Driver Genes.

Table [5] shows the confusion matrix of the binary-class problem and Table [6] shows the confusion matrix for the multi-class problem.

		Prediction	
		Cancer	**No Cancer**
Actual	**Cancer**	246	2
	No Cancer	4	18

Table 5 Confusion Matrix for Binary-class problem using Cancer Driver Genes

		Prediction				
		Head and Neck	**Lungs**	**Breast**	**Stomach**	**No Cancer**
Actual	**Head and Neck**	43	4	0	2	0
	Lungs	7	73	0	1	0
	Breast	0	2	65	0	0
	Stomach	1	2	0	45	1
	No Cancer	2	0	0	0	22

Table 6 Confusion Matrix for Multi-class problem using Cancer Driver Genes

The artificial neural network model was then trained on the 50 features extracted from each site using the feature extraction method as explained in section 3.4. The total number of features extracted in this set using the feature extraction method was 94 as some of the features were present in more than one site.

The artificial neural network model was trained and tested 5 times and average F score was calculated. The model was trained for 1500 epochs in each iteration. This was done for both multi-label and binary classifications. The F Score for multi-label classification was reported as 0.91 and F-Score for binary classification was reported 0.97 on the test set.

Figure [7] shows the training and validation accuracy for the multi-class problem and Figure [8] shows the training and validation accuracy for the binary-class problem for each epoch.

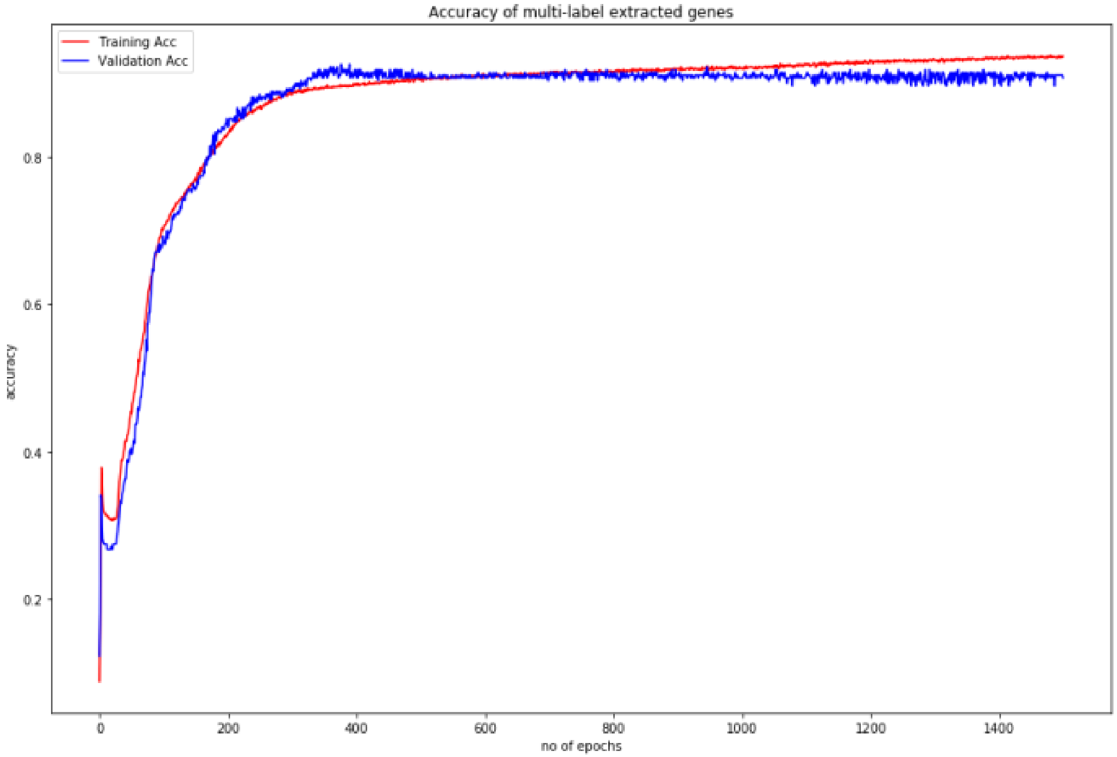

Figure 7 Training and Validation Accuracy for Multi-class problem using the Extracted Features.

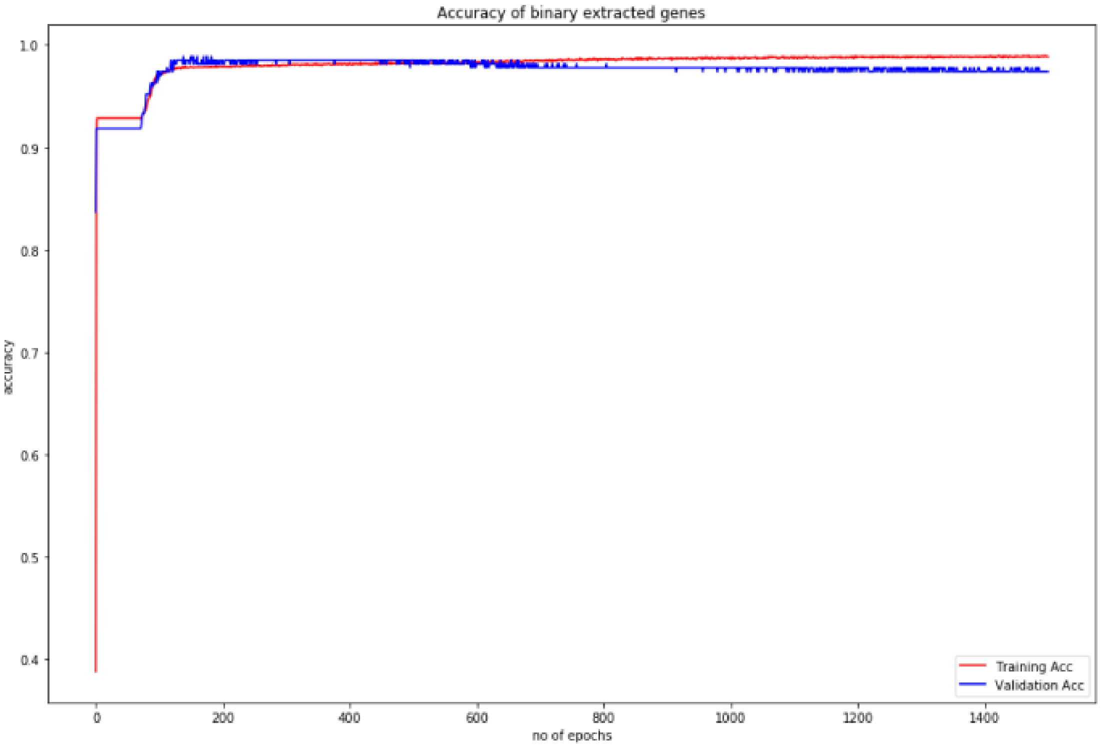

Figure 8 Training and Validation Accuracy for Binary-class problem using the Extracted Features.

Table [7] shows the confusion matrix of the binary-class problem and Table [9] shows the confusion matrix for the multi-class problem for the extracted features using the feature extraction method.

		Prediction	
		Cancer	No Cancer
Actual	Cancer	246	2
	No Cancer	5	17

Table 7 Confusion Matrix for Binary-class problem using the Extracted Features.

		Prediction				
		Head and Neck	Lungs	Breast	Stomach	No Cancer
Actual	Head and Neck	42	6	0	1	0
	Lungs	5	74	0	2	1
	Breast	0	1	65	1	1
	Stomach	0	2	0	46	1
	No Cancer	2	0	2	0	18

Table 8 Confusion Matrix for Multi-class problem using the Extracted Features.

By comparing the results of table [5] with table [7] we can see that in the case of cancer vs non cancer classification problem the mRMR feature extraction methods give approximately the same results as that to the model that was trained on the cancer driver genes data as the model trained on the extracted features only miss classified one non cancer sample as compared to the model that was trained on the cancer driver genes.

By comparing the results of table [6] with table [8] we can see that the same trend is followed by the multiclass problem as was in the binary classification problem. The mRMR feature extraction methods give approximately the same results as that to the model that was trained on the cancer driver genes data as the model trained on the extracted features only miss classified five samples as compared to the model that was trained on the cancer driver genes.

From the above finding we can conclude that the 50 features extracted from the mRMR feature extraction method can be used to classify between benign and malignant cancer in the 4 regions of the human body.

Figure [9] and figure [10] shows the F-score on the test set comparison of the different studies against the method proposed in this study.

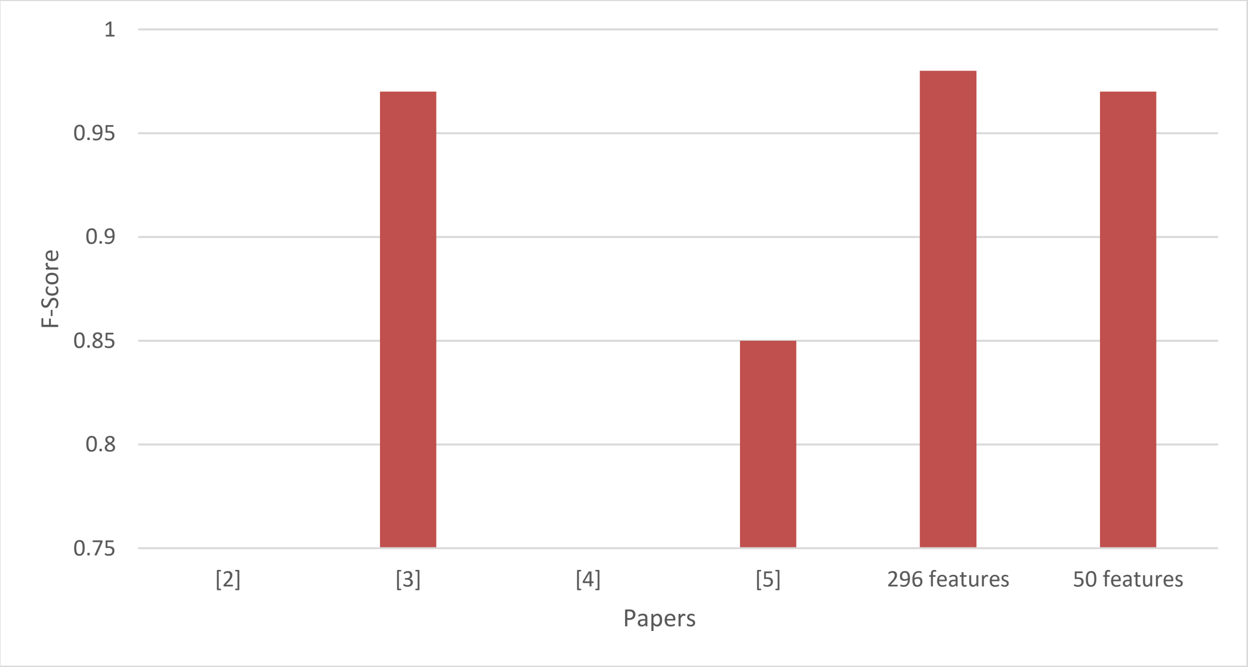

Figure 9 F-score comparison of binary classification problem of different studies

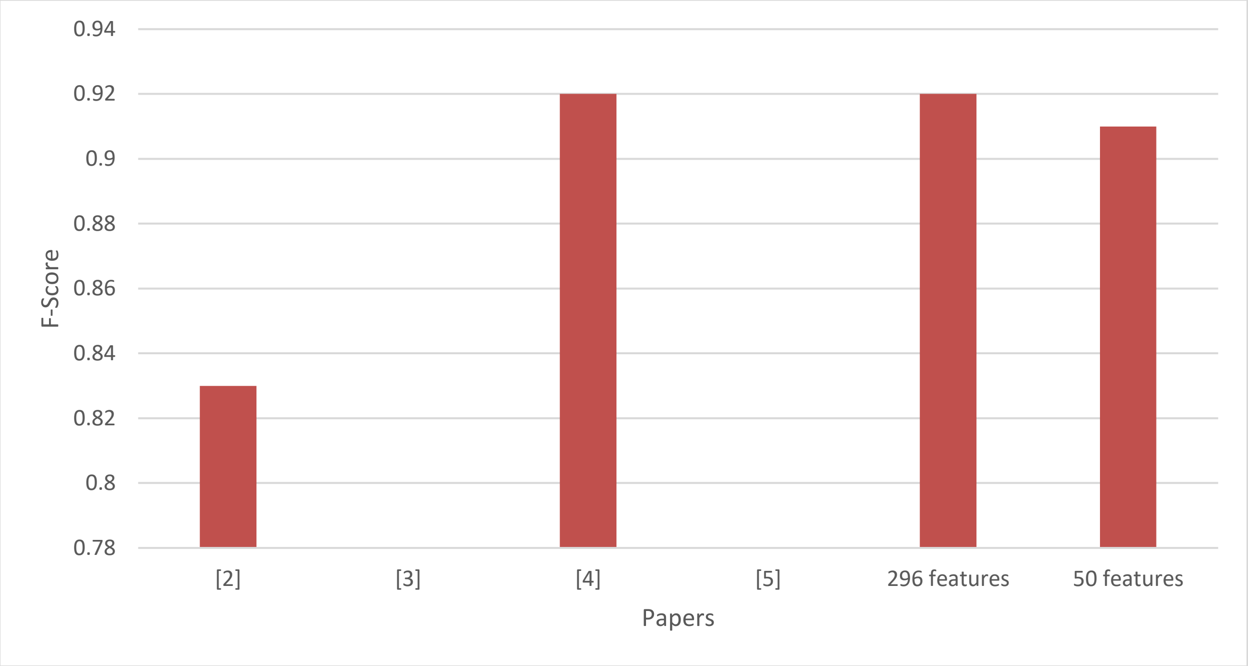

Figure 10 F-score comparison of multi classification problem of different studies

From figures [9] and [10] we can see that the F-scores of the methods proposed in this study are equal or greater than the F-scores of the previous studies in both the binary class problem and in the multi-class problem. In one of the instances in the multi-class problem where the F-score of the 50 features extracted model is lower than the F-score of one of study but the F-score is just 1 percent lower than the F-score value of the study. This shows that we can achieve comparable results to the other studies while using a very small feature set which in turn can be attributed to reduced space complexity.

Table [9] lists the hardware that was used in each study.

Parameters	Studies			
	[2]	**[3]**	**[4]**	**[5]**
CPU	no data available	Intel Xeon E5 2670 v2 @2.5Ghz	Intel i7	no data available
RAM		128 GB per node	12 GB	
Cores		20 per node	6	
Threads		**--**	12	
Nodes		66	--	

Table 9 System Specifications of each study

By comparing the data from Table [2] to that of Table [9] we can see that the system used in this study is too much underpower as compared to the systems used in other studies. One reason for using an underpowered system was that by utilizing the data preprocessing and feature extraction methods we were able to train and test the Artificial Neural Network model on a low-end system without compromising on the over-all accuracy of the model on the test set.

By using a small sub-set of the total features by employing knowledge of the prior studies and modeling a small but efficient neural network we were able to reduce the space complexity of the problem, which in turn reduced the time complexity of the problem as well.

Apart from reducing the space complexity of the problem, the feature extraction method also has another effect, that it can give an insight as to which genes are important in the cancer causation. This can be done by mapping the associated CpG site starting points of each extracted features to their associated genes present in the preprocessed data. Table [10] shows the genes associated with each site that were extracted using the feature extraction method. Table [10] shows that not all the genes in the cancer driver genes are present in the sites which shows that not all genes are responsible for cancer causation in the whole body.

Site	Gene Symbol	Number of Genes
Driver Genes	NOTCH1, MAP3K1, CBL, STAG2, KDM5C, JAK1, CDC73, IDH1, CIC, STK11, CARD11, PTPN11, GNAQ, MSH6, MPL, ALK, SMO, RUNX1, MET, SETBP1, HIST1H3B, BCL2, FOXL2, KLF4, CREBBP, EP300, EGFR, NRAS, ARID1A, TRAF7, DNMT3A, ACVR1B, JAK2, PTEN, GATA3, PIK3R1, NOTCH2, ABL1	38
Lungs	ALK, NOTCH1, NRAS, SMO, HIST1H3B, GATA3, TRAF7, CARD11, SETBP1, MAP3K1, GNAQ, STAG2, RUNX1, EGFR, ABL1, STK11, JAK1, FOXL2	18
Breast	NOTCH1, EP300, CARD11, GATA3, ALK, TRAF7, RUNX1, SETBP1, MAP3K1, GNAQ, JAK1, STAG2, STK11, EGFR, MPL, HIST1H3B, SMO, MET, FOXL2	19
Head and Neck	ALK, NOTCH1, NRAS, CARD11, HIST1H3B, GATA3, TRAF7, EP300, SETBP1, CREBBP, MAP3K1, GNAQ, JAK1, STAG2, EGFR, MPL, ARID1A, STK11, SMO, CIC, PIK3R1	21
Stomach	ALK, NOTCH1, NRAS, SMO, STK11, HIST1H3B, GATA3, TRAF7, SETBP1, CREBBP, JAK2, CARD11, RUNX1, PTEN, MAP3K1, JAK1, GNAQ	17

Table 10 Genes associated with extracted features of each site.

CHAPTER 5 CONCLUSION AND FUTURE WORK

5.1 Conclusion:

This study proposes a new method of classifying the benign and malignant tumors in 4 sites of the human body which cause the greatest number of deaths in the world, based on a data driven approach. First all those methylation sites were removed from the data that were associated with the passenger genes that do not contribute to the cancer causation. A small four layered ANN was trained and tested on the reduced feature-set which gave F-scores that were either matching the scores of previous studies or were beating them but never coming short in both the binary classification problem and in multi-class classification problem.

The study then proceeded to further reduce the number of features while keeping the F-score comparable. This was achieved using a feature extraction method along side the ANN and we were able to achieve similar F-scores in both the binary classification problem and multi-class classification problem while using a smaller feature-set.

One of the advantages of using a smaller feature-set was that we got an insight on the genes that are responsible in the cancer causation for the 4 regions that were selected. Also, we observed that not all the driver genes are active in all four regions showing that each site has a different set of genes that are responsible for the cancer causation.

Second advantage of using a smaller feature-set was that we reduced the overall space complexity of the problem. By reducing the space complexity of the problem, we were able to eliminate the need for a high specification and high-performance computer as was demonstrated in this study.

By using feature extraction method alongside an ANN, we were able to match the F-scores of previous studies while reducing the space complexity of the over-all problem.

5.2 Future Work:

The finding of this study can be extended to regions of the body as only four out of twenty-eight regions were selected in the study. Using the same feature extraction method along with the ANN region specific genes can be identified.

The study can then be further extended to find the genes responsible for the cancer causation in multiple sites of a single region giving further insights about the genes responsible for cancer.

Exact number of features can be selected using the hoping technique as described in section 3.4.2 by incorporating smaller jumps of one or two features between forty-five and fifty-five number of features.

The same ANN with a different feature extraction or the same feature extraction method can be used on the passenger genes to find if there are genes that are being missed just because they have been labeled as passenger genes.

APPENDIX A Data Preprocessing Source Code (MATLAB)

```
[~,~,X] = xlsread('file_names.xlsx');
O = xlsread('Oncogenes_TSGS.xlsx');
O = O(:,1);
b = 0 ;
[~,~,Y] = xlsread(Data.xlsx');
  start = Y(1,:);
  a = 1;
  for d =1: size(O,1)
    i = O(d,1);
      star(:,a) = Y(:,i); % start
      a = a+1;
  end
  star(:,a) = Y(:,size(Y,2));
  % removning rows with NA in beta values and -1 in start points
  c = 1;
  for d = 1: a-1
        done(c,1) = star(d);
      c = c+1;
  end
xlswrite(Preprocessed_Data.xlsx',star);
```

APPENDIX B Feature Extraction Source Code

```
import pandas as pd

import pymrmr


data = pd.read_csv(‘Preprocessed_Data.csv')

d = pymrmr.mRMR(data, 'MIQ', 50)

d.to_csv('mrmr features.csv')
```

APPENDIX C ANN Source Code

```
import numpy as np
import tensorflow as tf
from tensorflow import keras
import matplotlib.pyplot as plt
import pandas as pd
import sklearn
from keras.models import Sequential
from keras.layers import Dense, Activation
from tensorflow.keras.optimizers import RMSprop
from keras import regularizers
from sklearn.metrics import f1_score
from sklearn import metrics

x_train = pd.read_csv('x training.csv', low_memory=False, header = None)
y_train = pd.read_csv('y training.csv', low_memory=False, header = None)
x_test = pd.read_csv('x data testing.csv', low_memory=False, header = None)
y_test = pd.read_csv('y data testing.csv', low_memory=False, header = None)

x_Train = x_train.as_matrix()
y_Train = y_train.as_matrix()
x_Test = x_test.as_matrix()
y_Test = y_test.as_matrix()
```

```
    model = keras.Sequential([
        keras.layers.Dense(148, input_dim=x_Train.shape[1], activation = 'relu',
kernel_initializer = keras.initializers.glorot_uniform(seed=10)),
        keras.layers.Dense(74, activation = 'relu', kernel_initializer =
keras.initializers.glorot_uniform(seed=10)),
        keras.layers.Dense(37, activation = 'relu', kernel_initializer =
keras.initializers.glorot_uniform(seed=10)),
        keras.layers.Dense(5, activation = 'sigmoid', kernel_initializer =
keras.initializers.glorot_uniform(seed=10))
        ])


    opt = keras.optimizers.RMSprop(lr=0.00001, rho=0.9)
    model.compile(optimizer= opt ,loss = 'categorical_crossentropy', metrics=['accuracy'])
    M = model.fit(x_Train, y_Train, epochs = 1500, validation_data =(x_Test,y_Test))


    P = model.predict(x_Test)
    P = np.argmax(P, axis=1)
    L = np.argmax(y_Test, axis=1)
    f1 = f1_score(L, P,  average='micro')
    print('f1 score: ')
    print(f1)
    confusion = sklearn.metrics.confusion_matrix(L, P, labels=None, sample_weight=None)
    print(confusion)
    print(np.count_nonzero(np.abs(P-L)))
```

REFERENCES

[1] https://www.who.int/news-room/fact-sheets/detail/cancer

[2] Markus List, et al. Classification of Breast Cancer Subtypes by combining Gene Expression and DNA Methylation Data. Journal of Integrative Bioinformatics (2014)

[3] Celli, Fabrizio & Cumbo, Fabio & Weitschek, Emanuel. (2018). Classification of Large DNA Methylation Datasets for Identifying Cancer Drivers. Big Data Research. 10.1016/j.bdr.2018.02.005.

[4] Chatterjee, Soham & Iyer, Archana & Avva, Satya & Kollara, Abhai & Sankarasubbu, Malaikannan. (2018). Convolutional Neural Networks In Classifying Cancer Through DNA Methylation.

[5] Cai, Zhihua & Xu, Dong & Zhang, Qing & Zhang, Jiexia & Ngai, Sai Ming & Shao, Jianlin. (2014). Classification of lung cancer using ensemble-based feature selection and machine learning methods. Mol. BioSyst.. 11. 10.1039/C4MB00659C.

[6] https://www.cancer.net/navigating-cancer-care/cancer-basics/genetics/genetics-cancer

[7] Paziewska A, Dabrowska M, Goryca K, et al. DNA methylation status is more reliable than gene expression at detecting cancer in prostate biopsy. Br J Cancer. 2014 Aug 12;111(4):781–789. PubMedPMID: 24937670; PubMed Central PMCID: PMCPMC4134497.

[8] Vogelstein, B. et al. Cancer genome landscapes. Science 339, 1546–1558 (2013).

[9] Lin RK, Wang YC. Dysregulated transcriptional and post-translational control of DNA methyltransferases in cancer. Cell Biosci. 2014;4:46. PubMed PMID: 25949795; PubMed Central PMCID: PMC4422219.

[10] Du,P. et al. (2010) Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics, 11, 587

[11] Peng, Hanchuan & Long, Fuhui & Ding, Chris. (2005). Feature Selection Based On Mutual Information: Criteria of Max-Dependency,Max-Relevance, and Min-Redundancy. IEEE transactions on pattern analysis and machine intelligence. 27. 1226-38. 10.1109/TPAMI.2005.159.

[12] https://www.cancer.gov/tcga

Classification of Cancer using Epigenetic Markers

ORIGINALITY REPORT

7%	4%	4%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	doctorpenguin.com Internet Source	1%
2	Submitted to Universita' Telematica Internazionale Nettuno Student Paper	<1%
3	Mansour Sheikhan, Mahdi Bejani, Davood Gharavian. "Modular neural-SVM scheme for speech emotion recognition using ANOVA feature selection method", Neural Computing and Applications, 2012 Publication	<1%