REVA UNIVERSITY
Bengaluru, India

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

A PROJECT REPORT
ON

"CANCER SUBTYPE PREDICTION"

submitted in partial fulfilment of the requirement for the award of the degree of

BACHELOR OF TECHNOLOGY
IN
COMPUTER SCIENCE AND ENGINEERING

Submitted by

Dileep Kumar Reddy J K
R18CS511

Soham Kishor Misal
R17CS404

Veerendra Patil P
R18CS535

Under the guidance of

Dr. Nimrita Koul
Associate Professor
School of CSE
REVA University

MAY 2022

Rukmini Knowledge Park, Kattigenahalli, Yelahanka, Bengaluru - 560 064

www.reva.edu.in

DECLARATION

We, Mr. Dileep Kumar Reddy J K, Mr. Soham Kishor Misal, and Mr. Veerendra Patil P, students of Bachelor of Technology, belonging to School of CSE, REVA University, declare that this Project Report/ Dissertation entitled "Cancer Subtype Prediction" is the result of the project/ dissertation work done by us under the supervision of Dr. Nimrita Koul, Associate Professor, at School of CSE, REVA University.

We are submitting this Project Report/ Dissertation in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering by the REVA University, Bengaluru during the academic year 2021-2022.

We declare that this project report has been tested for plagiarism and has passed the plagiarism test with the similarity score less than 25% and it satisfies the academic requirements in respect of Project work prescribed for the said Degree.

We further declare that this project/ dissertation report or any part of it has not been submitted for award of any other Degree/ Diploma of this University or any other University/ Institution.

Signature of the candidates

Signed by us on

Certified that this project work submitted by Dileep Kumar Reddy J K, Soham Kishor Misal, and Veerendra Patil P has been carried out under my guidance and the declaration made by the candidates is true to the best of my knowledge.

Signature of Guide
Date:

Signature of Director/
Deputy Director of School
Date:

Official Seal of the School

REVA UNIVERSITY
Bengaluru, India

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

Certified that the project work entitled **Cancer Subtype Prediction** carried out under my guidance by **Dileep Kumar Reddy J K (R18CS511), Soham Kishor Misal (R17CS404) and Veerendra Patil P (R18CS535)**, bonafide students of REVA University during the academic year 2021-22, are submitting the project report in partial fulfillment for the award of **Bachelor of Technology** in Computer Science and Engineering during the academic year **2021–22**. The project report has been tested for plagiarism and has passed the plagiarism test with the similarity score less than 25%. The project report has been approved as it satisfies the academic requirements in respect of Project work prescribed for the said Degree.

Signature with Date

Dr. Nimrita Koul
(Guide)

Signature with Date

Dr. Ashwin Kumar U M
(Deputy Director)

Signature with Date

Dr. M Dhanamjaya
(Vice Chancellor)

External Examiner

Sl. No	Name of Examiner with Affiliation	Signature with Date
1.		
2.		

ACKNOWLEDGEMENT

Any given task that is achieved is never the result of the efforts of a single individual. There is always a group of people who play an instrumental role in leading a task to its completion. Our joy at having successfully completed the project would be incomplete without thanking everyone who helped us along the way. We would like to express our sincere gratitude to REVA University for providing the means for attaining this cherished goal.

We would like to thank Hon’ble Chancellor, **Dr. P Shyama Raju**, Hon’ble Vice-Chancellor, **Dr. M Dhanamjaya** for their immense support towards students to showcase their innovative ideas.

We would like to thank Deputy Director of School of Computer Science and Engineering, **Dr. Ashwin Kumar U M,** for providing us with a highly conductive environment and encouraging the growth and creativity of every student.

We would like to take this opportunity to express our gratitude towards our guide, **Dr. Nimrita Koul**, Associate Professor, School of Computer Science and Engineering for constantly supporting and guiding us throughout the course of the project.

Finally, we would like to thank the **Teaching and Non-Teaching Faculty** of REVA University for their unwavering support.

ABSTRACT

Cancer is caused as a result of unconstrained cell growth. It has several subtypes, identification of these subtypes in a quick and efficient manner is crucial in the treatment of cancer patients. In this paper the TCGA RNA-Seq dataset is chosen for training the Deep Learning based CNN model to predict the subtypes of cancer. Several pre-processing methods such as handling missing data, feature selection and normalization are applied. The feature selection technique used is Recursive Feature Elimination, it helps select 50 genes out of the available 20,531 genes. The gene data corresponding to each patient is stored in a NumPy array. The array is then used to create heat maps with the help of imshow() matplotlib function. The dataset contains 33 labels. The CNN model consists of 7 convolutional layers, each consisting of a kernel size of 3x3, 7 pooling layers, 7 batch normalization layers, 2 dense layers and 1 dropout layer. ReLu is the activation function used for the aforementioned layers. Softmax is the activation function used for the last dense layer. In order to avoid overfitting a dropout rate of 0.15 is used. The model provides a test accuracy of 73.87%.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1 Introduction

Cancer is a condition that begins with aberrant cell conduct and division, resulting in damage to neighboring cells and culminating in a lump or tumor, which can lead to death in some circumstances. It is ranked as the second biggest cause of death worldwide, accounting for one out of every six fatalities. Early detection and treatment can help to limit the risk of damage to neighboring cells. Therefore, to reduce the impact of cancer on people's health, significant research initiatives have been directed towards its screening and therapy strategies. The goal of cancer diagnosis is to classify tumors and identify indicators for each malignancy so that we may construct a learning system that can detect cancer early on. The need for implementing Artificial Intelligence to identify new genetic markers is becoming a crucial element in many biomedical applications, with heightened understanding of targeted therapy and timely identification strategies progressing over decades of technological advancements, accomplishing a responsiveness of around 80%.

RNA-Seq is a relatively recent and widely used approach for detecting novel isoforms and transcripts by giving additional normalized and far less imprecise data for diagnosis and detection. Identifying differentially expressed genes in the body or discovering gene changes at various levels is one of most essential function of transcriptome profiling. RNA-sequencing allows for simultaneous detection and characterization. Data of RNA-Seq is easily obtainable from several databases and it is being utilized to identify various cancer types. Moreover, due to their unprecedented proportion, complication, and the presence of repetitions in attribute values, RNA gene expression data analysis is particularly difficult. As a result, there is a demand for automatic feature extraction, that can be met using machine learning and deep learning techniques.

Deep learning is a new area based on recent advances in machine learning. It is a method that seeks to work on arriving at a decision derived from unprocessed data without taking into consideration the phases of extraction of features. It is why the phrase "automated feature engineering" was coined. Deep learning is currently being employed in a variety of fields. A convolutional neural network is a deep learning model for dealing with vast amounts of graphical data.

CNN captures the most significant features and decreases the neural network's complexity by making use of several approaches. Several illness detection techniques use deep learning, which is enhancing the performance of machine learning in the sector. A recent technology used in deep learning to recognise and classify various forms of tumors is a feed forward neural network also called as a multilayer perceptron (MLP).

The Cancer Genome Atlas (TCGA), which contains more than 11,000 tumors representing 33 of the most common types of cancer, is a well-known resource for cancer transcriptome profiling.

1.2 PROBLEM DEFINITION

To develop an approach and test it for accurate identification and prediction of the subtypes of various cancerous tumors.

1.3 OBJECTIVES

- To study the RNA-Sequence dataset
- To apply data pre-processing methods on the dataset
- To convert the dataset into images
- To build a CNN model to predict the subtypes of cancer

CHAPTER 2: LITERATURE SURVEY

For classifying pan-cancer, the authors of paper [1] have utilised the GA/KNN approach. The characteristic selection engine is the genetic algorithm (GA), and the algorithm used for classification is the k-nearest neighbours (KNN) method. They were able to uncover multiple groups of 20 genes which could properly categorise well over 90% of the data from 31 types of tumours in a validation dataset just by making use of the RNA-Seq expression of genes.

To help diagnose and evaluate cancer the authors of paper [2] made use of unsupervised feature learning with the help of data from gene expression. The key advantage of the suggested approach above earlier cancer detection systems is the ability to automatically create features from data from multiple forms of cancer to aid in its diagnosis of a particular type. To determine and identify cancer, the system provides a more thorough and generic strategy.

The authors of paper [3] have made use of the TCGA RNA-Seq data to categorize 30+ various types of cancer patients. They compared the efficiency, learning period, accuracy, recalls, and F1-scores of 5 machine learning methods, namely decision tree (DT), k nearest neighbour (KNN), linear support vector machine (linear SVM), polynomial support vector machine (poly SVM), and artificial neural network (ANN). The results demonstrate that linear SVM is the top classifier in the investigation, with an overall accuracy of 95.8%.

The researchers of paper [4] used TCGA RNA-Seq data from about 30 various types of cancer patients, as well as healthy tissue RNA-Seq data from GTEx. One thousand and twenty-four genes with the greatest up or down regulation counts across the entire dataset are chosen. The input for model training is the expression data of the selected genes. The training data is converted to RGB colours by transforming gene expression levels into binary format of 24 bits.

A Convolutional Neural Network (CNN) model is used to carry out the training of the model. The proposed algorithm has an accuracy of 97%.

The authors of paper [5] created a model based on deep learning that uses 3 diverse layers of information to distinguish pan-cancer metastatic status. The model was created with data of four hundred patients from TCGA. They quantified the suggested convolutional variational autoencoder and alternative feature extraction approaches demonstrating that using mRNA, microRNA, and DNA methylation data as attributes improved the performance of their model when compared to simply using mRNA data. Furthermore, they demonstrated that mRNA-related traits played a larger role in computationally distinguishing initial tumours from metastatic tumours. Finally, their deep learning model surpassed a machine learning ensemble approach on a variety of criterias.

The authors of paper [6] suggested that if classification mistakes are managed, then the study of unconstrained tissue elements of cancer and determining pan-cancer subgroups could be addressed by utilising tissue related molecular markers. They proposed that when the PAM50, a commercially popular and accessible cancer hallmark is combined with unknown evaluation, it can be remodelled for a pan-cancer setting, resulting in multiple groups having therapeutic, biological, and molecular consequences.

Using large volume of RNA-Seq and scRNA-Seq data, the authors of paper [7] developed cancer predictors that can recognise twenty-one kinds of tumours and normal tissues. Relying just on 300 highly relevant genes present in each tumour, the system was trained with nearly seven thousand cancer samples and around six hundred normal samples from twenty-one malignancies and normal tissues present in the TCGA dataset. They then compared the outputs of various machine learning algorithms with Artificial Neural Network. The Artificial Neural Network regularly outperformed the other approaches. They next implemented their method to scRNA-Seq data that had been smoothed with kNN and discovered that the system accurately categorised cancer kinds and normal samples.

In the first step, the authors of paper [8] used a component significance ranking scheme to select several key genes. They then used a good classifier to assess the categorization ability of all simple combinations of such essential genes.

Their approach achieved an extremely high accuracy with only 2 or 3 genes for 3 data sets each containing 2, 3, and 4 types of cancer. They separated the problem into a series of dual categorization problems and performed the two-step strategy to all these dual categorization problems for a big and complicated dataset containing fourteen kinds of cancer.

The authors of paper [9] have proposed 2 new descriptions of multiclass relevant attributes. One of the attributes Full Class Relevant stands for possible biomarkers that can be used to distinguish between different cancer kinds. Partial Class Relevant genes, on the other hand, identifies subsets of different cancers. They've presented a Markov blanket embedded memetic method for identifying both FCR and PCR genes at the same time. The suggested method corresponds to legitimate FCR and PCR genes that will aid researchers in their study, according to findings acquired on regularly used artificial and authentic microarray sets of data. On several datasets of microarray, it has been discovered that identifying both FCR and PCR genes improves accuracy rate.

CHAPTER 3: FEASIBILITY STUDY

3.1 Compliance with Society, Ethical and Social Practices

Cancer is caused as a result of unconstrained cell growth, resulting in damage to neighbouring cells and culminating in a lump or tumour, which can lead to death in some circumstances. To reduce the impact of cancer on people's health, significant research initiatives have been directed towards its screening and therapy strategies. The goal of cancer diagnosis is to classify tumours and identify indicators for each malignancy so that we may construct a learning system that can detect cancer early on. Doctors fail to identify cancer in 10 to 28 percentage of patients, artificial intelligence overcomes this problem by detecting minor patterns that humans overlook.

The deep learning based Convolutional Neural Network (CNN) model used in this project helps diagnose the subtypes of cancer. The model will benefit medical personals by providing them with precise and accurate diagnosis of cancer subtypes and help save their valuable time. The CNN model is non-discriminatory and treats all users equally.

3.2 Compliance with Environment and Legal Feasibility

- The CNN model proposed in this project is a software application which is efficient in performance, as a result of this the runtime of the model is significantly less. Short runtime of the model leads to low carbon footprint from the machine running it.
- The project is feasible legally as the dataset used to train the CNN model is open source.

CHAPTER 4: METHODOLOGY

The proposed method uses a deep learning based Convolutional Neural Network to predict the subtypes of cancer.

The structure of the proposed system is shown in Figure.1

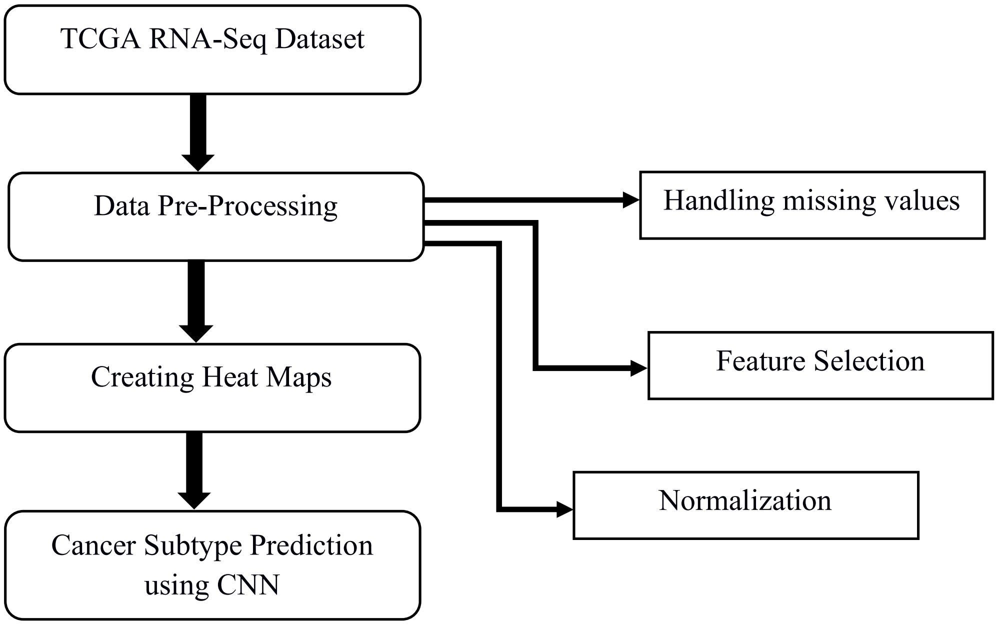
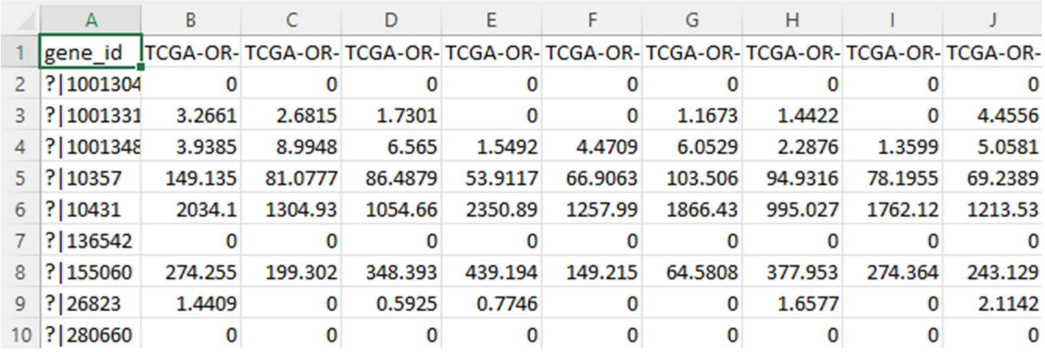


Figure.1 Diagram of proposed system

(i). Dataset:

The dataset used in this project is the TCGA RNA-Seq dataset.



	A	B	C	D	E	F	G	H	I	J
1	gene_id	TCGA-OR-	TCGA-OR-	TCGA-OR-	TCGA-OR-	TCGA-OR-	TCGA-OR-	TCGA-OR-	TCGA-OR-	TCGA-OR-
2	?|1001304	0	0	0	0	0	0	0	0	0
3	?|1001331	3.2661	2.6815	1.7301	0	0	1.1673	1.4422	0	4.4556
4	?|1001348	3.9385	8.9948	6.565	1.5492	4.4709	6.0529	2.2876	1.3599	5.0581
5	?|10357	149.135	81.0777	86.4879	53.9117	66.9063	103.506	94.9316	78.1955	69.2389
6	?|10431	2034.1	1304.93	1054.66	2350.89	1257.99	1866.43	995.027	1762.12	1213.53
7	?|136542	0	0	0	0	0	0	0	0	0
8	?|155060	274.255	199.302	348.393	439.194	149.215	64.5808	377.953	274.364	243.129
9	?|26823	1.4409	0	0.5925	0.7746	0	0	1.6577	0	2.1142
10	?|280660	0	0	0	0	0	0	0	0	0

Figure. 2. Screenshot of a part of the TCGA RNA-Seq dataset

(ii). Pre-processing:

The process of converting raw data into a comprehensible format is known as data pre-processing. Some of the pre-processing methods used are:

a. Missing Data:

When no data value is maintained for a variable in an observation, missing values emerge. Missing data is ubiquitous, and it can have a big impact on the inferences that can be taken from the data. Therefore, the null values present in the dataset are dropped by making use of the pandas dropna() method.

b. Feature Selection:

When creating a predictive model, feature selection is the method of minimising the number of parameters. The quantity of input variables should be reduced to lower the cost of computation and increase the model's performance. The feature selection method used in this project is the Recursive Feature Elimination. Recursive Feature Elimination is an attribute selection approach that eliminates the lowest attribute/ attributes up until the set of attributes provided is achieved.

Recursive Feature Elimination technique is applied on the TCGA RNA-Seq dataset to select 50 genes out of the available 20,531 genes.

c. Normalization:

It is the process of converting data so that it appears on the same scale across all elements in a dataset.

The 50 selected genes are normalized in the range 0 to 255.

(iii). Heat Maps:

It is a 2D information visualisation approach that depicts the intensity of an event as colour. In order to create heat maps, the data present in the csv file is first transposed. Now the patient ids are represented in rows and the various types of genes are represented in columns. The gene values of each patient are fed to a NumPy array. The matplotlib function imshow() is used to create images from the 2-dimensional NumPy arrays.

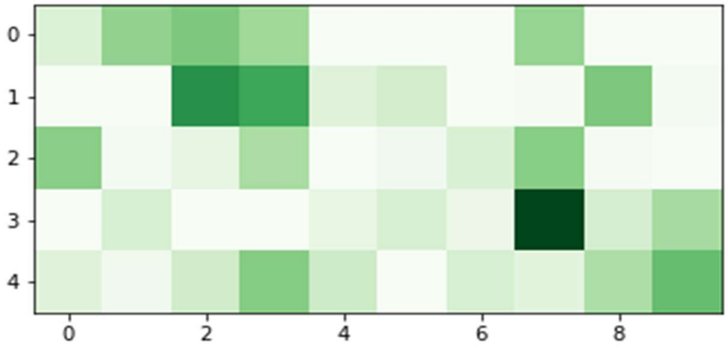
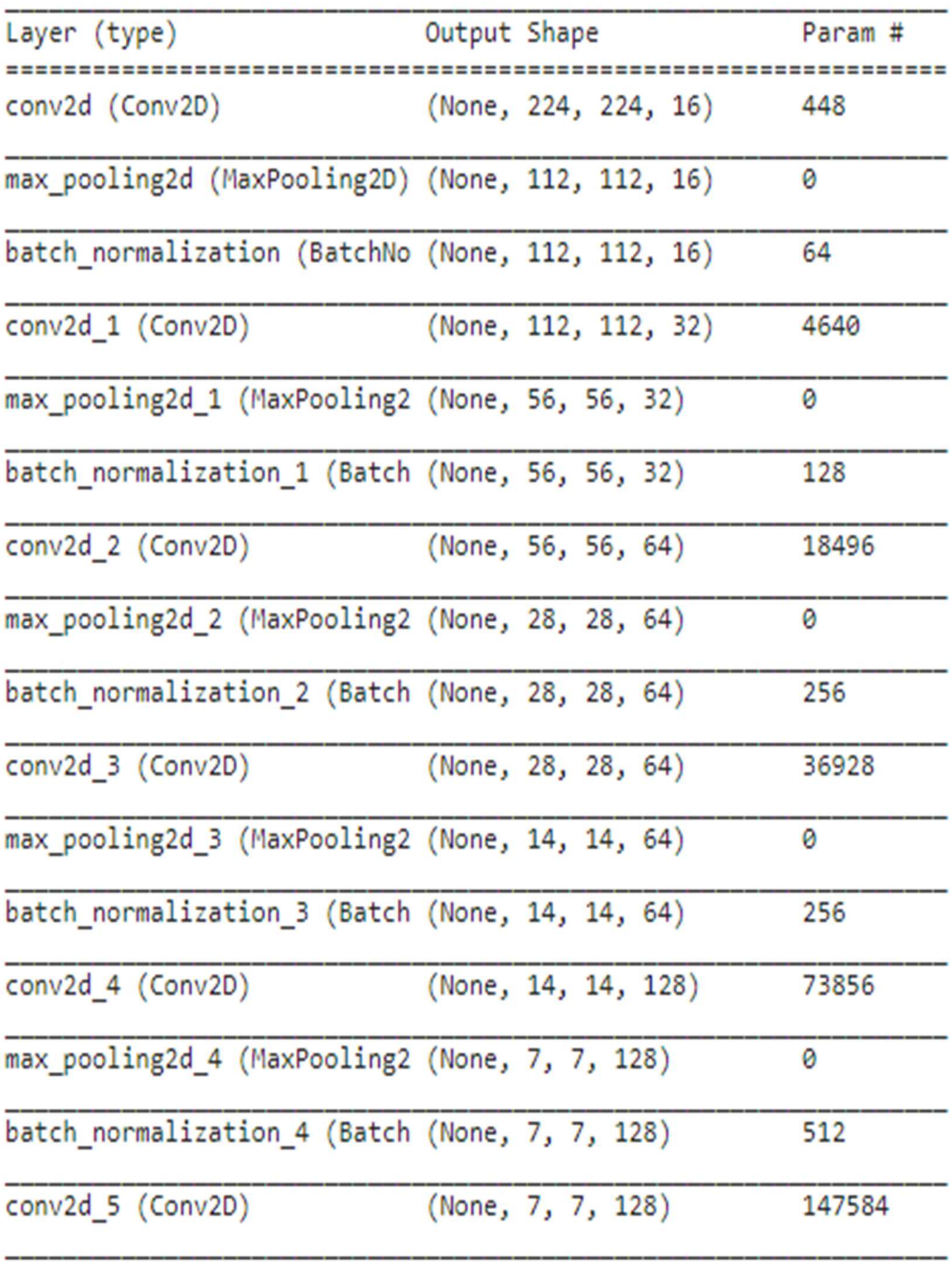


Figure. 3 Heat Map of cancer type ACC

(iv). Model Architecture:

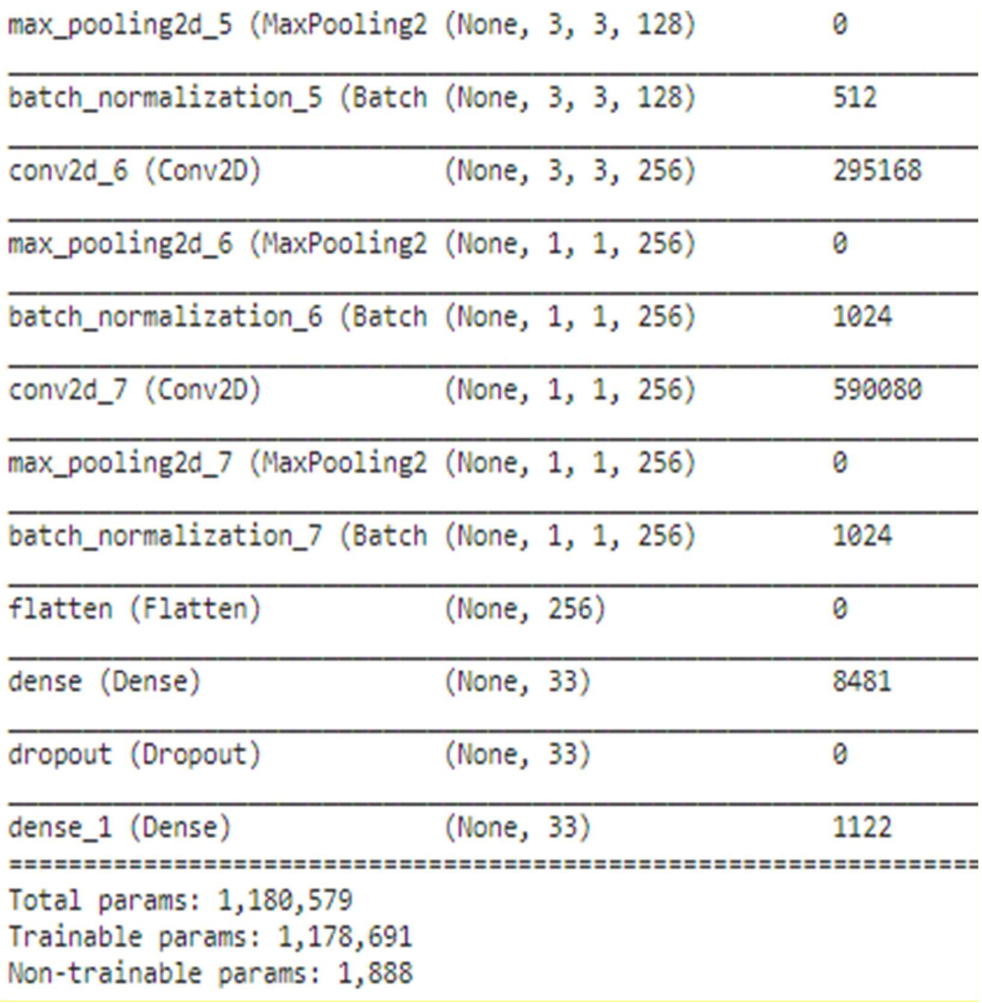
The CNN architecture represented by Figure.4.1 and Figure 4.2 is used for training, it consists of 7 convolutional layers each consisting of a kernel size of 3x3, 7 pooling layers, 7 batch normalization layers, 2 dense layers and 1 dropout layer. ReLu is the activation function used for the aforementioned layers.

Softmax is the activation function used for the last dense layer. In order to avoid overfitting, the dropout rate of 0.15 is used.



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 224, 224, 16)      448
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 112, 112, 16)      0
_________________________________________________________________
batch_normalization (BatchNo (None, 112, 112, 16)      64
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 112, 112, 32)      4640
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 56, 56, 32)        0
_________________________________________________________________
batch_normalization_1 (Batch (None, 56, 56, 32)        128
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 56, 56, 64)        18496
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 28, 28, 64)        0
_________________________________________________________________
batch_normalization_2 (Batch (None, 28, 28, 64)        256
_________________________________________________________________
conv2d_3 (Conv2D)            (None, 28, 28, 64)        36928
_________________________________________________________________
max_pooling2d_3 (MaxPooling2 (None, 14, 14, 64)        0
_________________________________________________________________
batch_normalization_3 (Batch (None, 14, 14, 64)        256
_________________________________________________________________
conv2d_4 (Conv2D)            (None, 14, 14, 128)       73856
_________________________________________________________________
max_pooling2d_4 (MaxPooling2 (None, 7, 7, 128)         0
_________________________________________________________________
batch_normalization_4 (Batch (None, 7, 7, 128)         512
_________________________________________________________________
conv2d_5 (Conv2D)            (None, 7, 7, 128)         147584
_________________________________________________________________
```

Figure. 4.1 CNN Model Architecture



```
max_pooling2d_5 (MaxPooling2 (None, 3, 3, 128)        0
_________________________________________________________________
batch_normalization_5 (Batch (None, 3, 3, 128)        512
_________________________________________________________________
conv2d_6 (Conv2D)            (None, 3, 3, 256)        295168
_________________________________________________________________
max_pooling2d_6 (MaxPooling2 (None, 1, 1, 256)        0
_________________________________________________________________
batch_normalization_6 (Batch (None, 1, 1, 256)        1024
_________________________________________________________________
conv2d_7 (Conv2D)            (None, 1, 1, 256)        590080
_________________________________________________________________
max_pooling2d_7 (MaxPooling2 (None, 1, 1, 256)        0
_________________________________________________________________
batch_normalization_7 (Batch (None, 1, 1, 256)        1024
_________________________________________________________________
flatten (Flatten)            (None, 256)              0
_________________________________________________________________
dense (Dense)                (None, 33)               8481
_________________________________________________________________
dropout (Dropout)            (None, 33)               0
_________________________________________________________________
dense_1 (Dense)              (None, 33)               1122
=================================================================
Total params: 1,180,579
Trainable params: 1,178,691
Non-trainable params: 1,888
```

Figure. 4.2 CNN Model Architecture

(v). Training:

The heat map images generated were of the order 432*288 pixels, before starting the training of the model they were reduced to 244*244 pixels. The CNN model makes use of 3,084 samples from 33 labels of tumors. The samples are split in the ratio of 20:80 for testing and training respectively.

CHAPTER 5: RESULTS

Performance:

The accuracy of the model is 73.87% after 50 epochs.

The accuracy & loss charts for the test and training data are displayed in Figure. 5. The accuracy, precision, recall, F1-Score and Cohen Kappa Score are shown in Figure. 6. The precision, recall and F1-Score for each of the 33 cancer classes are given in Figure. 7. The overall accuracy of the model is given in Figure. 8.

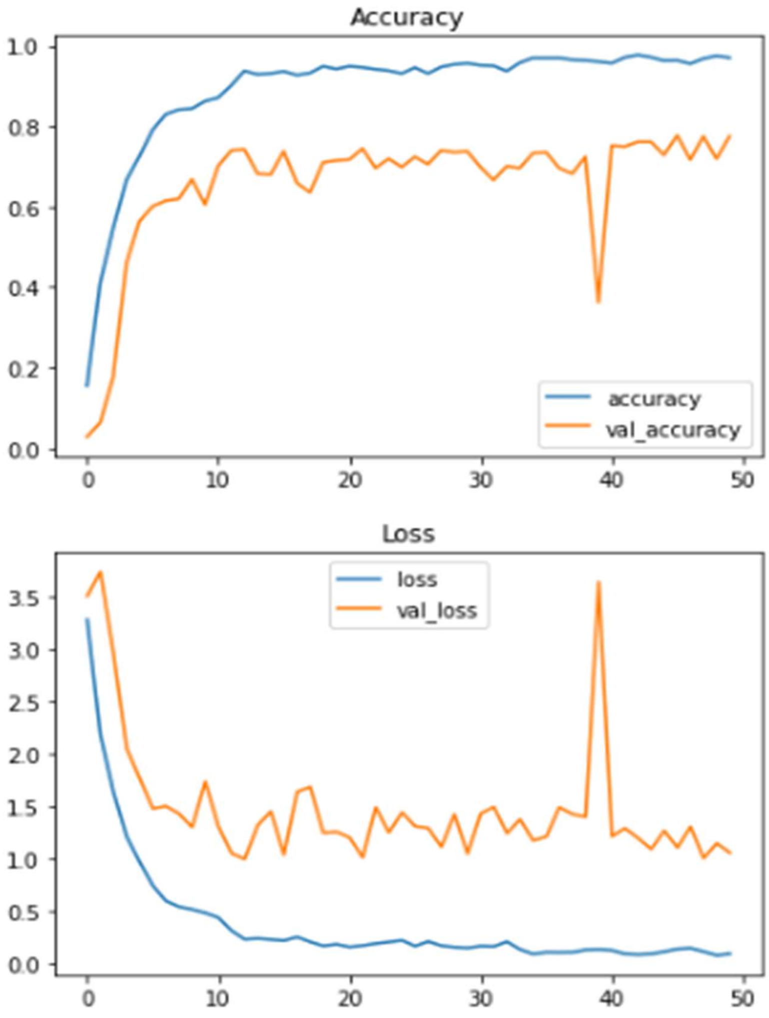
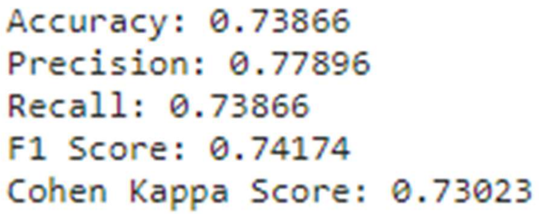


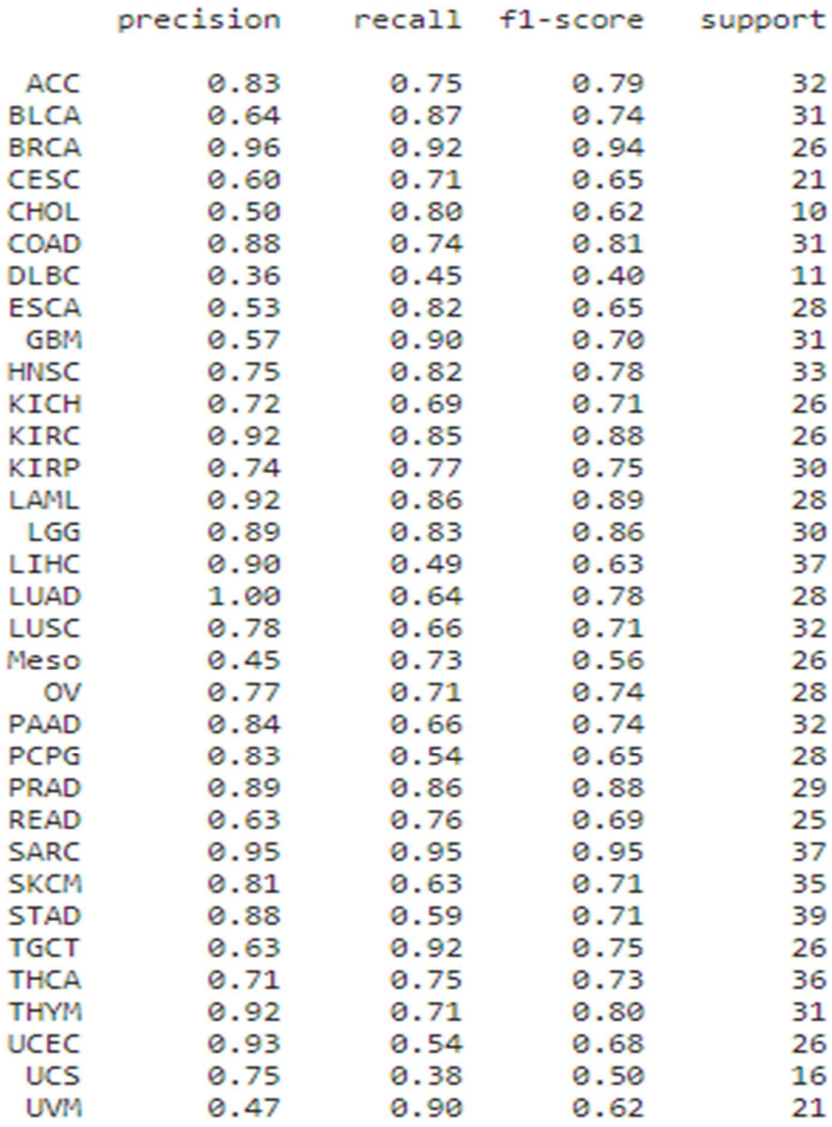
Figure. 5 Accuracy and Loss charts for test and training data

Note: Blue represents test data and orange represents training data



```
Accuracy: 0.73866
Precision: 0.77896
Recall: 0.73866
F1 Score: 0.74174
Cohen Kappa Score: 0.73023
```

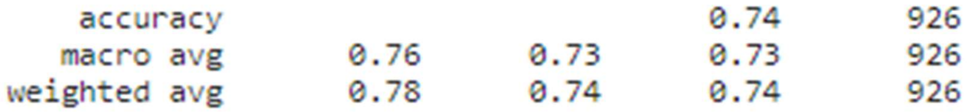
Figure. 6 Accuracy, Precision, Recall, F1 Score & Cohen Kappa Score



```
              precision    recall  f1-score   support

         ACC       0.83      0.75      0.79        32
        BLCA       0.64      0.87      0.74        31
        BRCA       0.96      0.92      0.94        26
        CESC       0.60      0.71      0.65        21
        CHOL       0.50      0.80      0.62        10
        COAD       0.88      0.74      0.81        31
        DLBC       0.36      0.45      0.40        11
        ESCA       0.53      0.82      0.65        28
         GBM       0.57      0.90      0.70        31
        HNSC       0.75      0.82      0.78        33
        KICH       0.72      0.69      0.71        26
        KIRC       0.92      0.85      0.88        26
        KIRP       0.74      0.77      0.75        30
        LAML       0.92      0.86      0.89        28
         LGG       0.89      0.83      0.86        30
        LIHC       0.90      0.49      0.63        37
        LUAD       1.00      0.64      0.78        28
        LUSC       0.78      0.66      0.71        32
        Meso       0.45      0.73      0.56        26
          OV       0.77      0.71      0.74        28
        PAAD       0.84      0.66      0.74        32
        PCPG       0.83      0.54      0.65        28
        PRAD       0.89      0.86      0.88        29
        READ       0.63      0.76      0.69        25
        SARC       0.95      0.95      0.95        37
        SKCM       0.81      0.63      0.71        35
        STAD       0.88      0.59      0.71        39
        TGCT       0.63      0.92      0.75        26
        THCA       0.71      0.75      0.73        36
        THYM       0.92      0.71      0.80        31
        UCEC       0.93      0.54      0.68        26
         UCS       0.75      0.38      0.50        16
         UVM       0.47      0.90      0.62        21
```

Figure. 7 Precision, Recall and F1-Score for each of the 33 cancer classes



```
    accuracy                           0.74       926
   macro avg       0.76      0.73      0.73       926
weighted avg       0.78      0.74      0.74       926
```

Figure. 8 Overall accuracy of the model

CHAPTER 6: COST ESTIMATION

There is no explicit cost except for the cost of publication, since the datasets and tools that are used in the project are open source.

CHAPTER 7: CONCLUSION

Cancer has several subtypes, identification of these subtypes in a quick and efficient manner is crucial in the treatment of cancer patients**.** The deep learning based CNN model that has been implemented in this project has been tested on the TCGA RNA-Seq dataset. This method provides a test accuracy of 73.87% on the multiclass dataset.

CHAPTER 8: FUTURE SCOPE

A front-end system can be implemented to accept images which will be provided as input to the CNN model. The model will predict the subtype of cancer and render the output to the user.

REFERENCES

[1] Li, Y., Kang, K., Krahn, J. M., Croutwater, N., Lee, K., Umbach, D. M., & Li, L. (2017). A comprehensive genomic pan-cancer classification using The Cancer Genome Atlas gene expression data. *BMC genomics*, *18*(1), *1-13.*

[2] Rasool Fakoor, Faisal Ladhak, Azade Nazi, Manfred Huber. (2013). Using deep learning to enhance cancer diagnosis and classification. *JMLR: W&CP volume 28.*

[3] Yi-Hsin Hsu, Dong Si. (2018). Cancer Type Prediction and Classification Based on RNA-sequencing Data. *PMID: 30441551.*

[4] Büşra Nur Darendeli, Alper Yılmaz. (2021). Convolutional Neural Network Approach to Predict Tumor Samples Using Gene Expression Data. *Journal of Intelligent Systems Theory and Applications, Volume 4, Issue 2, 136-141, 23.09.21.*

[5] Albaradei, S., Napolitano, F., Thafar, M. A., Gojobori, T., Essack, M., & Gao, X. (2021). MetaCancer: a deep learning-based pan-cancer metastasis prediction model developed using multiomics data. *Computational and Structural Biotechnology Journal*, *19*, 4404-4411.

[6] Rocha, D., García, I. A., González Montoro, A., Llera, A., Prato, L., Girotti, M. R., & Fernández, E. A. (2021). Pan-Cancer Molecular Patterns and Biological Implications Associated with a Tumor-Specific Molecular Signature. *Cells*, *10*(1), 45.

[7] Kim, B. H., Yu, K., & Lee, P. C. (2020). Cancer classification of single-cell gene expression data by neural network. *Bioinformatics*, *36*(5), 1360-1366.

[8] Wang L, Chu F, Xie W, "Accurate cancer classification using expressions of very few genes", *IEEE Transactions on Computational Biology and Bioinformatics, vol. 4, no. 1, 2007, pp. 40–53.*

[9] Zexuan Zhu, Y. S. Ong and M. Zurada, Identification of full and partial class relevant genes, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *vol. 7, no. 2, pp. 263-277, 2010.*

TURNITIN PLAGIARISM REPORT

ORIGINALITY REPORT

17%	6%	6%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to University of East London Student Paper	5%
2	Submitted to Liberty University Student Paper	1%
3	Nour Eldeen M. Khalifa, Mohamed Hamed N. Taha, Dalia Ezzat Ali, Adam Slowik, Aboul Ella Hassanien. "Artificial Intelligence Technique for Gene Expression by Tumor RNA-Seq Data: A Novel Optimized Deep Learning Approach", IEEE Access, 2020 Publication	1%
4	Submitted to Nottingham Trent University Student Paper	1%
5	doctorpenguin.com Internet Source	1%
6	encyclopedia.pub Internet Source	1%
7	Peijun Lu, Ning Gao, Zhaohua Lu, Jingjing Yang, Ou Bai, Qi Li. "Combined CNN and LSTM for Motor Imagery Classification", 2019 12th	1%

	International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 2019 Publication	
8	repository.aust.edu.ng Internet Source	1 %
9	Lipo Wang, Yaoli Wang, Qing Chang. "Feature selection methods for big data bioinformatics: A survey from the search perspective", Methods, 2016 Publication	1 %
10	Submitted to Liverpool John Moores University Student Paper	1 %
11	dergipark.org.tr Internet Source	1 %
12	Somayah Albaradei, Francesco Napolitano, Maha A. Thafar, Takashi Gojobori, Magbubah Essack, Xin Gao. "MetaCancer: A deep learning-based pan-cancer metastasis prediction model developed using multi-omics data", Computational and Structural Biotechnology Journal, 2021 Publication	1 %
13	Submitted to University of St Andrews Student Paper	<1 %
	Www.mdpi.com	

14	Internet Source	<1 %
15	keep.lib.asu.edu Internet Source	<1 %
16	Yi-Hsin Hsu, Dong Si. "Cancer Type Prediction and Classification Based on RNA-sequencing Data", 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2018 Publication	<1 %
17	link.springer.com Internet Source	<1 %
18	www.mdpi.com Internet Source	<1 %
19	vinayakumar ravi, Soman KP, Mamoun Alazab, Sriram S, Simran k. "A Comprehensive Tutorial and Survey of Applications of Deep Learning for Cyber Security", Institute of Electrical and Electronics Engineers (IEEE), 2020 Publication	<1 %

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		

From: IACIT 2022
Sent: 14 May 2022 11:01
To: Soham Kishor Misal
Cc: Dr.Nimrita Koul
Subject: IACIT-2022 | Acceptance Notification for Paper Id - 670

Dear Author/Authors,

Greetings from IACIT-2022 | School of Computer Science and Engineering, REVA University, Bengaluru.

Congratulations!!

We would like to inform you that your paper is accepted for presentation in **"4th International Virtual Conference on Advances in Computing and Information Technology (IACIT-2022)"** and publication in the Scopus Indexed Journal.

Paper Id: 670

Paper Title: Cancer Subtype Prediction

Note:
Publication charge for Scopus indexed journal is **14,000-00 INR (Fourteen Thousand rupees)**.

REVA University is sponsoring 50% of the publication charges for REVA – CSE STUDENTS.

Kindly pay SEVEN Thousand rupees (7000 INR) for Scopus Journal.

Account Details:

Beneficiary Name	: FACE
A/C Numbers	: 6662500101063001
IFSC code	: KARB0000666
Branch	: REVA University
Bank	: KARNATAKA BANK

The deadline for Registration **is 14th May, 2022**

Feel free to contact Prof.K V Sheelavathy-9901492266 for any queries.

You are requested to fill the Google form once you are done with the payment.

Cancer Subtype Prediction

Soham Kishor Misal
Computer Science & Engineering
REVA University
Bengaluru, India
r17cs404@cit.reva.edu.in

Dileep Kumar Reddy J K
Computer Science & Engineering
REVA University
Bengaluru, India
r18cs511@cit.reva.edu.in

Veerendra Patil P
Computer Science & Engineering
REVA University
Bengaluru, India
r18cs535@cit.reva.edu.in

Dr. Nimrita Koul
Associate Professor
Computer Science & Engineering
REVA University
Bengaluru, India
nimrita.koul@reva.edu.in

Abstract* – Cancer is caused as a result of unconstrained cell growth. It has several subtypes, identification of these subtypes in a quick and efficient manner is crucial in the treatment of cancer patients. TCGA RNA-Seq dataset is chosen for training the Deep Learning Model. Several pre-processing methods such as handling missing data, feature selection and normalization are applied. The feature selection technique used is Recursive Feature Elimination, it helps select 50 genes out of 20,531. The gene data corresponding to each patient is stored in a NumPy array. The array is then used to create heat maps with the help of imshow() matplotlib function. The dataset contains 33 labels. A CNN model is built to predict the subtype of cancer. The model has an accuracy of 73.87%.

***Keywords* – Cancer, Convolutional Neural Network (CNN), Deep Learning (DL), Recursive Feature Elimination (RFE), TCGA, RNA-Seq**

I. INTRODUCTION

Cancer is ranked as the second biggest cause of death worldwide, accounting for one out of every six fatalities. To reduce the impact of cancer on people's health, significant research initiatives have been directed towards its screening and therapy strategies. The goal of cancer diagnosis is to classify tumors and identify indicators [1, 2, 3] for each malignancy so that we may construct a learning system that can detect cancer early on. The need for implementing Artificial Intelligence to identify new genetic markers is becoming a crucial element in many biomedical applications, with heightened understanding of targeted therapy and timely identification strategies progressing over decades of technological advancements, accomplishing a responsiveness of around 80%. The Cancer Genome Atlas (TCGA) [11], which contains more than 11,000 tumors representing 33 of the most common types of cancer, is a well-known resource for cancer transcriptome profiling.

II. RELATED WORK

For classifying pan-cancer, the authors of paper [1] have utilised the GA/KNN approach.
The characteristic selection engine is the genetic algorithm (GA), and the algorithm used for classification is the k-nearest neighbours (KNN) method. They were able to uncover multiple groups of 20 genes which could properly categorise well over 90% of the data from 31 types of tumours in a validation dataset just by making use of the RNA-Seq expression of genes.

To help diagnose and evaluate cancer the authors of paper [2] made use of unsupervised feature learning [5, 6] with the help of data from gene expression [7, 8]. The key advantage of the suggested approach above earlier cancer detection systems is the ability to automatically create features from data from multiple forms of cancer to aid in its diagnosis of a particular type. To determine and identify cancer, the system provides a more thorough and generic strategy.

The authors of paper [3] have made use of the TCGA RNA-Seq data [11] to categorize 30+ various types of cancer patients. They compared the efficiency, learning period, accuracy, recalls, and F1-scores of 5 machine learning methods, namely decision tree (DT), k nearest neighbour (KNN), linear support vector machine (linear SVM), polynomial support vector machine (poly SVM), and artificial neural network (ANN). The results demonstrate that linear SVM [9, 10] is the top classifier in the investigation, with an overall accuracy of 95.8%.

The researchers of paper [4] used TCGA RNA-Seq data [11] from about 30 various types of cancer patients, as well as healthy tissue RNA-Seq data from GTEx. One thousand and twenty four genes with the greatest up or down regulation counts across the entire dataset are chosen. The input for model training is the expression data of the selected genes.

The training data is converted to RGB colours by transforming gene expression levels into binary format of 24 bits. A Convolutional Neural Network (CNN) model is used to carry out the training of the model. The proposed algorithm has an accuracy of 97%.

III. DATASET

The TCGA RNA-Seq dataset is chosen to train the CNN model, it contains 33 different types of cancer, they are ACC, BLCA, BRCA, CESC, CHOL, COAD, DLBC, ESCA, GBM, HNSC, KICH, KIRC, KIRP, LAML, LGG, LIHC, LUAD, LUSC, MESO, OV, PAAD, PCPG, PRAD, READ, SARC, SKCM, STAD, TGCT, THCA, THYM, UCEC, UCS and UVM.

IV. METHODS

(i). Pre-processing:

a. Missing Data:

The null values present in the dataset are dropped by making use of the pandas dropna() method.

b. Feature Selection:

Recursive Feature Elimination technique is applied to select 50 genes out of the available 20,531 genes.

c. Normalization:

The 50 selected genes are normalized in the range 0 to 255.

(ii). Heat Maps:

In order to create heat maps, the data present in the csv file is first transposed. Now the patient ids are represented in rows and the various types of genes are represented in columns. The gene values of each patient are fed to a NumPy array. The matplotlib function imshow() is used to create images from the 2-dimensional NumPy arrays.

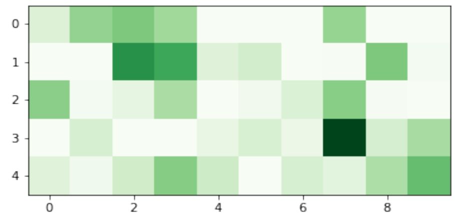
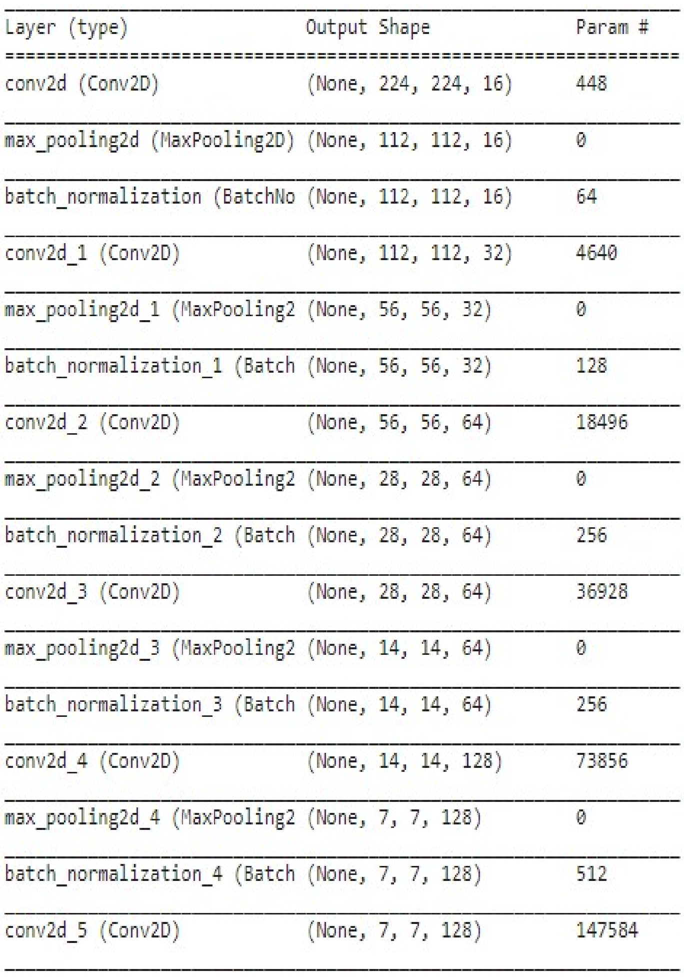


Figure. 1 Heat Map of cancer type ACC

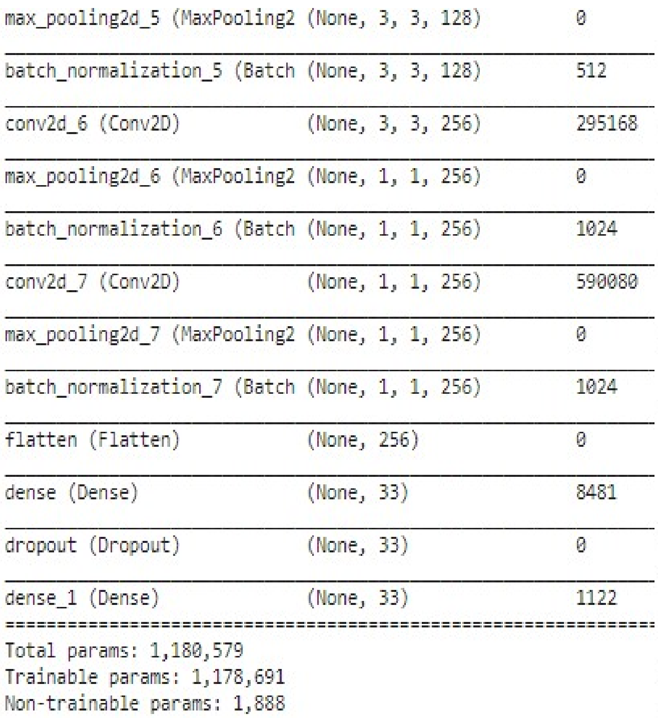
(iii). Model Architecture:

The CNN architecture represented by Figure.2.1 and Figure. 2.2 is used for training, it consists of 7 convolutional layers each consisting of a kernel size of 3x3, 7 pooling layers, 7 batch normalization layers, 2 dense layers and 1 dropout layer. ReLu is the activation function used for the aforementioned layers. Softmax is the activation function used for the last dense layer. In order to avoid overfitting, the dropout rate of 0.15 is used.



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 224, 224, 16)      448
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 112, 112, 16)      0
_________________________________________________________________
batch_normalization (BatchNo (None, 112, 112, 16)      64
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 112, 112, 32)      4640
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 56, 56, 32)        0
_________________________________________________________________
batch_normalization_1 (Batch (None, 56, 56, 32)        128
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 56, 56, 64)        18496
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 28, 28, 64)        0
_________________________________________________________________
batch_normalization_2 (Batch (None, 28, 28, 64)        256
_________________________________________________________________
conv2d_3 (Conv2D)            (None, 28, 28, 64)        36928
_________________________________________________________________
max_pooling2d_3 (MaxPooling2 (None, 14, 14, 64)        0
_________________________________________________________________
batch_normalization_3 (Batch (None, 14, 14, 64)        256
_________________________________________________________________
conv2d_4 (Conv2D)            (None, 14, 14, 128)       73856
_________________________________________________________________
max_pooling2d_4 (MaxPooling2 (None, 7, 7, 128)         0
_________________________________________________________________
batch_normalization_4 (Batch (None, 7, 7, 128)         512
_________________________________________________________________
conv2d_5 (Conv2D)            (None, 7, 7, 128)         147584
_________________________________________________________________
```

Figure. 2.1 Architecture of CNN Model



```
max_pooling2d_5 (MaxPooling2 (None, 3, 3, 128)         0
_________________________________________________________________
batch_normalization_5 (Batch (None, 3, 3, 128)         512
_________________________________________________________________
conv2d_6 (Conv2D)            (None, 3, 3, 256)         295168
_________________________________________________________________
max_pooling2d_6 (MaxPooling2 (None, 1, 1, 256)         0
_________________________________________________________________
batch_normalization_6 (Batch (None, 1, 1, 256)         1024
_________________________________________________________________
conv2d_7 (Conv2D)            (None, 1, 1, 256)         590080
_________________________________________________________________
max_pooling2d_7 (MaxPooling2 (None, 1, 1, 256)         0
_________________________________________________________________
batch_normalization_7 (Batch (None, 1, 1, 256)         1024
_________________________________________________________________
flatten (Flatten)            (None, 256)               0
_________________________________________________________________
dense (Dense)                (None, 33)                8481
_________________________________________________________________
dropout (Dropout)            (None, 33)                0
_________________________________________________________________
dense_1 (Dense)              (None, 33)                1122
=================================================================
Total params: 1,180,579
Trainable params: 1,178,691
Non-trainable params: 1,888
```

Figure. 2.2 Architecture of CNN Model

(iv). Training:

The heat map images generated were of the order 432*288 pixels, before starting the training of the model they were reduced to 244*244 pixels. The CNN model makes use of 3,084 samples from 33 labels of tumors. The samples are split in the ratio of 20:80 for testing and training respectively.

(v). Performance:

The accuracy of the model is 73.87% after 50 epochs. The accuracy & loss charts for the test and training data are displayed in Figure. 3. The accuracy, precision, recall, F1-Score and Cohen Kappa Score are shown in Figure. 4. The precision, recall and F1-Score for each of the 33 cancer classes are given in Figure. 5. The overall accuracy of the model is given in Figure. 6. The confusion matrix is given in Figure. 7.

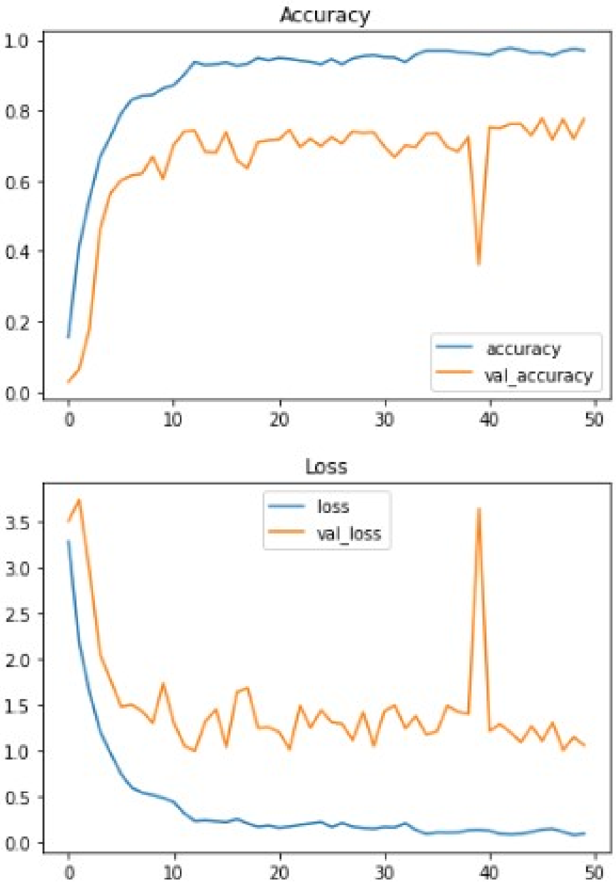
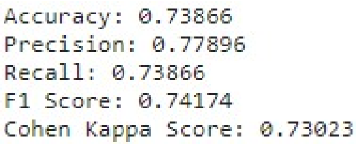


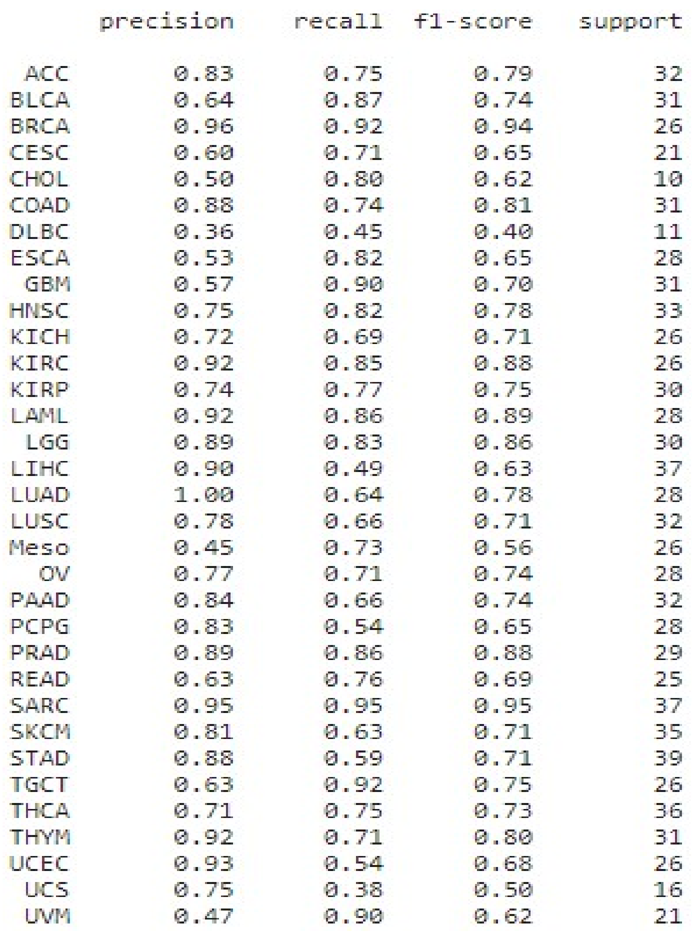
Figure. 3 Accuracy and Loss charts for test and training data

Note: Blue represents test data and orange represents training data.



```
Accuracy: 0.73866
Precision: 0.77896
Recall: 0.73866
F1 Score: 0.74174
Cohen Kappa Score: 0.73023
```

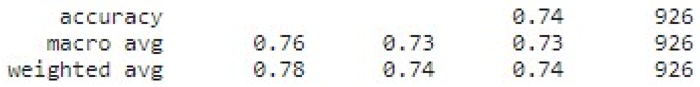
Figure. 4 Accuracy, Precision, Recall, F1 Score & Cohen Kappa Score



```
       precision    recall  f1-score   support

 ACC        0.83      0.75      0.79        32
BLCA        0.64      0.87      0.74        31
BRCA        0.96      0.92      0.94        26
CESC        0.60      0.71      0.65        21
CHOL        0.50      0.80      0.62        10
COAD        0.88      0.74      0.81        31
DLBC        0.36      0.45      0.40        11
ESCA        0.53      0.82      0.65        28
 GBM        0.57      0.90      0.70        31
HNSC        0.75      0.82      0.78        33
KICH        0.72      0.69      0.71        26
KIRC        0.92      0.85      0.88        26
KIRP        0.74      0.77      0.75        30
LAML        0.92      0.86      0.89        28
 LGG        0.89      0.83      0.86        30
LIHC        0.90      0.49      0.63        37
LUAD        1.00      0.64      0.78        28
LUSC        0.78      0.66      0.71        32
Meso        0.45      0.73      0.56        26
  OV        0.77      0.71      0.74        28
PAAD        0.84      0.66      0.74        32
PCPG        0.83      0.54      0.65        28
PRAD        0.89      0.86      0.88        29
READ        0.63      0.76      0.69        25
SARC        0.95      0.95      0.95        37
SKCM        0.81      0.63      0.71        35
STAD        0.88      0.59      0.71        39
TGCT        0.63      0.92      0.75        26
THCA        0.71      0.75      0.73        36
THYM        0.92      0.71      0.80        31
UCEC        0.93      0.54      0.68        26
 UCS        0.75      0.38      0.50        16
 UVM        0.47      0.90      0.62        21
```

Figure. 5 Precision, Recall and F1-Score for each of the 33 cancer classes



```
    accuracy                           0.74       926
   macro avg       0.76      0.73      0.73       926
weighted avg       0.78      0.74      0.74       926
```

Figure. 6 Overall accuracy of the model

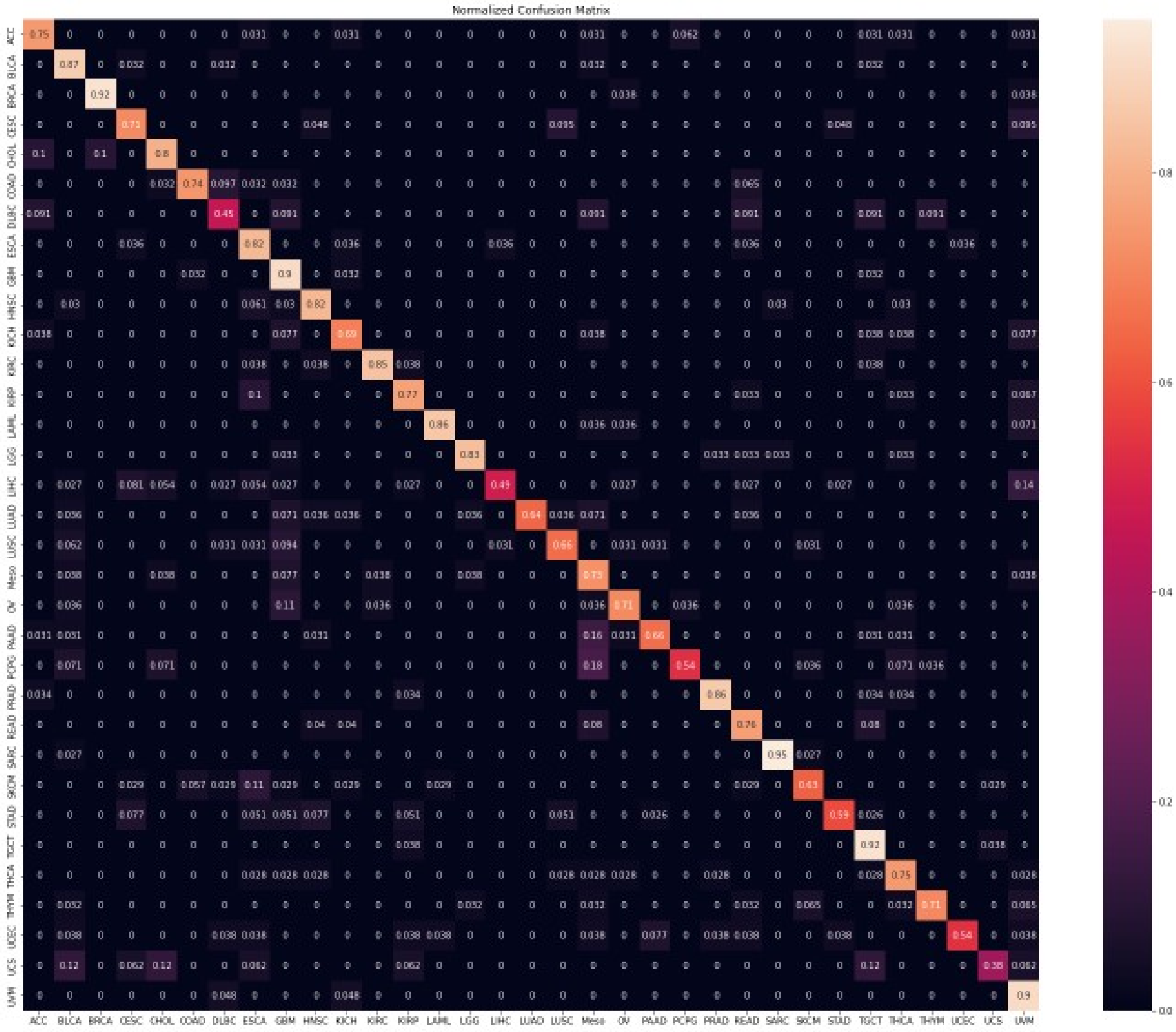


Figure. 7 Confusion Matrix

V. RESULT

Accuracy of the CNN Model is 73.87%.

VI. CONCLUSION

Cancer has several subtypes, identification of these subtypes in a quick and efficient manner is crucial in the treatment of cancer patients**.** The deep learning based CNN model that has been implemented in this paper has been tested on the TCGA RNA-Seq dataset. This method provides a test accuracy of 73.87% on this multiclass dataset.

VII. ACKNOWLEDGEMENT

The authors would like to thank the Department of Science and Technology, Government of India for supporting this research with the grant DSTICPS 2018.

VIII. REFERENCES

1. Li, Y., Kang, K., Krahn, J. M., Croutwater, N., Lee, K., Umbach, D. M., & Li, L. (2017). A comprehensive genomic pan-cancer classification using The Cancer Genome Atlas gene expression data. *BMC genomics*, *18*(1), *1-13.*

2. Rasool Fakoor, Faisal Ladhak, Azade Nazi, Manfred Huber. (2013). Using deep learning to enhance cancer diagnosis and classification. *JMLR: W&CP volume 28.*

3. Yi-Hsin Hsu, Dong Si. (2018). Cancer Type Prediction and Classification Based on RNA-sequencing Data. *PMID: 30441551.*

4. Büşra Nur Darendeli, Alper Yılmaz. (2021) Convolutional Neural Network Approach to Predict Tumor Samples Using Gene Expression Data. *Journal of Intelligent Systems Theory and Applications, Volume 4, Issue 2, 136-141, 23.09.21.*

5. Wang L, Chu F, Xie W, "Accurate cancer classification using expressions of very few genes", *IEEE Transactions on Computational Biology and Bioinformatics, vol. 4, no. 1, 2007, pp. 40–53.*

6. Zexuan Zhu, Y. S. Ong and M. Zurada, Identification of full and partial class relevant genes, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *vol. 7, no. 2, pp. 263-277, 2010.*

7. Mohammed Loey, Mohammed Wajeeh Jasim, Hazem M. EL-Bakry, Mohamed Hamed N. Taha, Nour Eldeen M. Khalifa "Breast and Colon Cancer Classification from Gene Expression Profiles Using Data Mining Techniques", *Symmetry vol. 12, no. 408, 2020, doi:10.3390/sym12030408*

8. M. A. H. Akhand, Md. Asaduzzaman Miah, Mir Hussain Kabir, M. M. Hafizur Rahman, Cancer Classification from DNA Microarray Data using mRMR and Artificial Neural Network, *International Journal of Advanced Computer Science and Applications, vol. 10, no. 7, 2019.*

9. Nada Almugren, Hala Alshamlana, "Survey on Hybrid Feature Selection Methods in Microarray Gene Expression Data for Cancer Classification",
IEEE Access, vol. 7, 2019 pp. 75833-44 10.1109/ACCESS.2019.2922987

10. Zakariya Yahya Algamal, Muhammad Hisyam Lee, "A two-stage sparse logistic regression for optimal gene selection in high-dimensional microarray data classification",
Advances in Data Analysis and Classification, vol. 13, pp:753–771, 2019

11. TCGA Dataset:
https://www.nature.com/articles/ng.2764

Acknowledgement Letter

It is to acknowledge that the Patent titled as "…………………………………………… …… ……………………..." has been received for filing.

The details of Applicants are:

First Applicant:

Second Applicant:

Third Applicant:

Fourth Applicant:

Fifth Applicant:

Sixth Applicant:

Filed Status (Mention Application Number): ……………………………………..

IPR Coordinator, SoCSE