Cancer Subtype Prediction

Dileep Kumar Reddy J K
Computer Science & Engineering
REVA University
Bengaluru, India
r18cs[illegible]@cit.reva.edu.in

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

Soham Kishor Misal
Computer Science & Engineering
REVA University
Bengaluru, India
r17cs404@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.

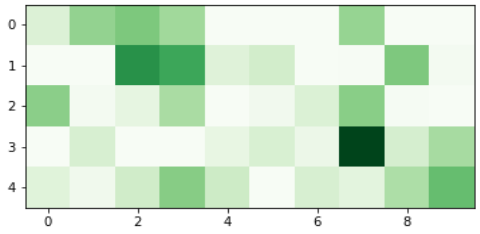

Figure. 1 Heat Map of cancer type ACC

(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.

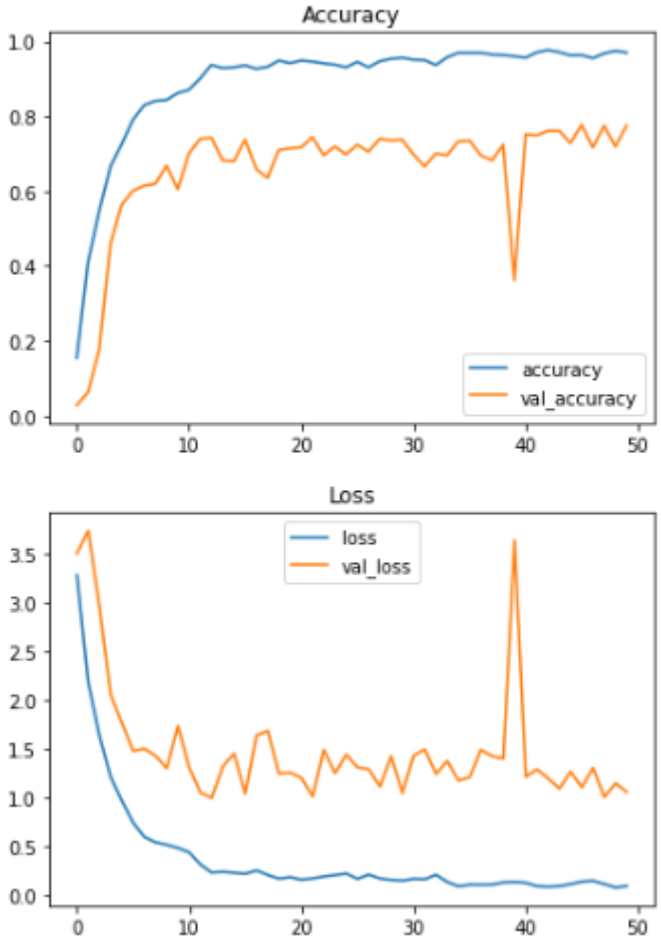

Figure. 3 Accuracy and Loss charts for test and training data
Note: Blue represents test data and orange represents training data.

```
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

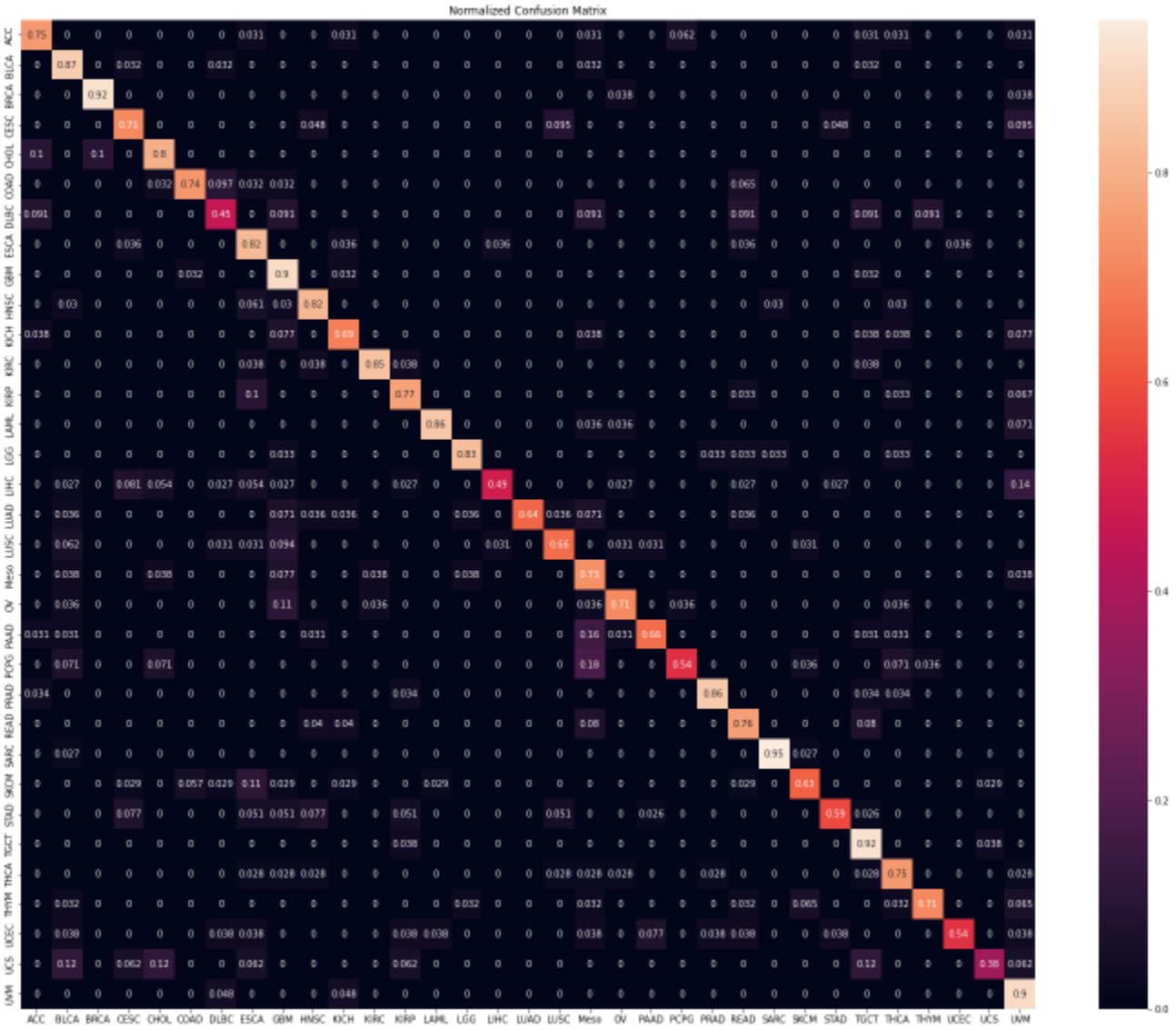

Figure. 7 Confusion Matri

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. Khalif "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