

**School of Computing Science and Engineering**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Course Code:** | | | **Major Project Progress Report – Phase 2** | | | **Year: 2021-2022** | | |
| **Sem: 8th** | | |
| **Project Group Members Details** | | | | | | | | |
| **Group No:** | | **Group Name:** | | | | | | |
| **Sl. No** | **SRN** | **Full Name** | | **Sec** | **Mobile No.** | **Email\_ID** | | **Sign** |
| **1** | R18CS511 | Dileep Kumar Reddy J K | | D | 8660589490 | [r18cs511@cit.reva.edu.in](mailto:r18cs511@cit.reva.edu.in) | |  |
| **2** | R17CS404 | Soham Kishor Misal | | D | 8310878092 | [r17cs404@cit.reva.edu.in](file:///D:\Drive\0.%207th\8.%20Major%20Project\0.0000%20P2\Report,%20Presentation%20final\r17cs404@cit.reva.edu.in) | |  |
| **3** | R18CS535 | Veerendra Patil P | | D | 9108747909 | [r18cs535@cit.reva.edu.in](file:///D:\Drive\0.%207th\8.%20Major%20Project\0.0000%20P2\Report,%20Presentation%20final\r18cs535@cit.reva.edu.in) | |  |
| **Project Details** | | | | | | | | |
| **Project Title:** | | **Generative Adversarial Network for Identification of Cancer Subtypes** | | | | | | |
| **Guide Details:** | | **Name: Dr. Nimrita Koul**  **Designation: Associate Professor** | | | | | **Mobile No:**  **8861097464** | |
| **Place of Project Work:** | |  | | | | | | |
| **External Guide:** | | **Name:**  **Designation:** | | | | | **Mobile No:** | |
| **Remarks by Guide:** | | | | | | | **Guide Signature**  **Date:** | |

***I. 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%.*

**II. 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 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), 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.  
  
**III. Problem Statement:**To develop an approach and test it for accurate identification and prediction of the subtypes of various cancerous tumors.

**IV. Literature Survey:**

|  |  |
| --- | --- |
| **Reference Number** | **Abstract** |
| [1] | For classifying pan-cancer, the authors 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. |
| [2] | The authors 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 [9, 10] is the top classifier in the investigation, with an overall accuracy of 95.8%. |
| [3] | The researchers of paper 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%. |
| [4] | The authors presented a deep learning-based model, that differentiates pan-cancer metastasis status based on three heterogeneous data layers. The model was built using 400 patients’ data that includes RNA sequencing, microRNA sequencing, and DNA methylation data from The Cancer Genome Atlas (TCGA). They have quantitatively assessed the proposed convolutional variational autoencoder (CVAE) and alternative feature extraction methods and have showed that by integrating mRNA, microRNA, and DNA methylation data as features they were able to improve their model's performance compared to when they used mRNA data only. In addition, they also showed that the mRNA-related features made a more significant contribution when attempting to distinguish the primary tumors from metastatic ones computationally. Lastly, their DL model significantly outperformed a machine learning (ML) ensemble method based on various metrics. |
| [5] | The authors of this paper suggested that studying tissue-independent components of cancer and defining pan-cancer subtypes could be addressed using tissue-specific molecular signatures if classification errors are controlled. They suggested that the PAM50 a well-known commercially available cancer signature could be repurposed for a pan-cancer context when paired with uncertainty assessment, resulting in two classes with molecular, biological, and clinical implications. |
| [6] | The authors designed cancer classifiers that can identify 21 types of cancers and normal tissues based on bulk RNA-Seq as well as scRNA-seq data. The training was performed with 7398 cancer samples and 640 normal samples from 21 tumors and normal tissues in TCGA based on the 300 most significant genes expressed in each cancer. They then compared neural network (NN), support vector machine (SVM), k-nearest neighbors (kNN) and random forest (RF) methods. The NN performed consistently better than other methods. They further applied their approach to scRNA-seq transformed by kNN smoothing and found that their model successfully classified cancer types and normal samples. |
| [7] | The authors in their first step, have chosen some important genes using a feature importance ranking scheme. In the second step, they have tested the classification capability of all simple combinations of those important genes by using a good classifier. For three "small" and "simple" data sets with two, three, and four cancer (sub)types, their approach obtained very high accuracy with only two or three genes. For a "large" and "complex" data set with 14 cancer types, they divided the whole problem into a group of binary classification problems and applied the 2-step approach to each of these binary classification problems. Through this "divide-and-conquer" approach, they obtained accuracy comparable to previously reported results but with only 28 genes rather than 16,063 genes. In general, their method can significantly reduce the number of genes required for highly reliable diagnosis. |
| [8] | In this paper the authors have we introduce two new definitions of multiclass relevancy features, i.e., full class relevant (FCR) and partial class relevant (PCR) features. FCR denotes genes that serve as candidate biomarkers for discriminating all cancer types. While PCR are genes that distinguish subsets of cancer types. They have proposed a Markov blanket embedded memetic algorithm for the simultaneous identification of both FCR and PCR genes. Results obtained on commonly used synthetic and real-world microarray data sets show that the proposed approach converges to valid FCR and PCR genes that would assist biologists in their research work. The identification of both FCR and PCR genes is found to generate improvement in classification accuracy on many microarray data sets. |

**V. Objectives:**

1. Study of the RNA-Sequence dataset

2. Data Pre-Processing

3. Conversion of RNA-Sequence dataset into images

4. Model building

5. Testing

**VI. Cost Estimation:**There is no explicit cost since the datasets and tools that are used in the project are open source.

**VII. Methodology:**

(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:

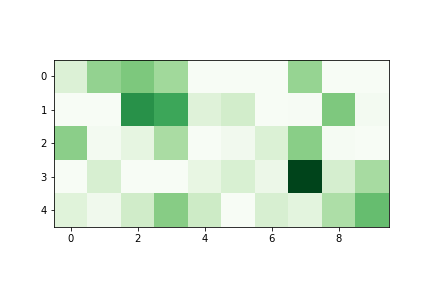
Recursive Feature Elimination technique is applied to select 50 genes out of the available 20,531 genes

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 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.

Table

Description automatically generated

Table

Description automatically generated

***Figure.2*** *CNN Model Architecture*

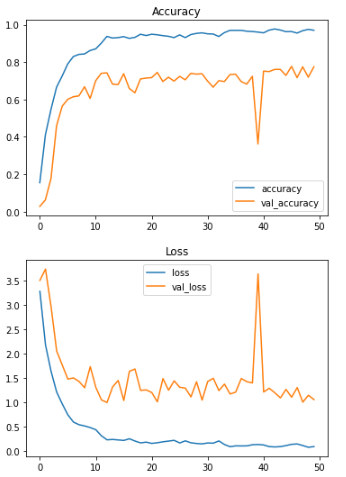
(iv). Training:

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.

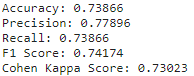
**VIII. Experimental Results:**

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. 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.*



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

Table

Description automatically generated

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

A screenshot of a computer

Description automatically generated with medium confidence

***Figure. 6*** *Overall accuracy of the model*

Chart

Description automatically generated

***Figure. 7*** *Confusion Matrix*

**IX. Duration Analysis:**

Phase 1:

|  |  |  |
| --- | --- | --- |
| Duration | Work | Status |
| October - Week 1 | Project idea discussion | Completed |
| October - Week 2 | Literature Survey | Completed |
| November - Week 1 | Identification of Modules | Completed |
| November - Week 2 | Concepts of Deep Learning | Completed |
| December - Week 1 | Defining Roles & Responsibilities | Completed |

Phase 2:

|  |  |  |
| --- | --- | --- |
| Duration | Work | Status |
| March - Week 3 | Study of RNA-Sequence Dataset | Completed |
| March - Week 4 | Data Pre-Processing | Completed |
| April - Week 2 | Model Building & Testing | Completed |
| April - Week 4 | Writing a Technical Paper | Completed |

**X. Compliance with Society, Ethical & Social Practices:**

The application will be available free of cost.

The model helps save time by detecting the type of cancer in a quick and efficient manner.

**XI. Compliance with Environment and Legal Feasibility:**

**XII. Future Enhancement:**

A front-end system can be implemented to accept images and predict the subtype of cancer.

**XIII. Conclusions:**

Cancer has several subtypes, identification of these subtypes in a quick and efficient manner is crucial in the treatment of cancer patients**.** The Generative Adversarial Network (GAN) based deep learning model built using TCGA RNA-Seq dataset will help identify the subtypes of cancer in a precise manner, thereby helping medical service providers save their valuable time.

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

[3] 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.*

[4] 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.

[5] 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.

[6] 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.

[7] 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.*

[8] 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.*