

**School of Computing Science and Engineering**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Course Code:** | | | **Major Project Progress Report – Phase 1** | | | **Year: 2020-2021** | | |
| **Sem: 7th Phase-I** | | |
| **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: Nimrita Koul**  **Designation: Assistant Professor** | | | | | **Mobile No:**  **8861097464** | |
| **Place of Project Work:** | |  | | | | | | |
| **External Guide:** | | **Name:**  **Designation:** | | | | | **Mobile No:** | |
| **Remarks by Guide:** | | | | | | | **Guide Signature**  **Date:** | |

***Abstract:***

*Cancer has been characterized as a heterogeneous disease consisting of several subtypes, prediction of these subtypes is vital for cancer diagnosis and therapy.   
The Cancer Genome Atlas (TCGA) has generated comprehensive molecular profiles for more than 30 different human cancer types.*

*Our aim is to build a deep learning model using Generative Adversarial Networks (GAN) on the TCGA RNA-Seq dataset to detect the types and subtypes of cancer and help medical service providers save time.*

1. **Introduction:**

Cancer is the second leading cause of death worldwide, an average of one in six deaths is due to cancer. Considerable research efforts have been devoted to cancer diagnosis and treatment techniques to lessen its impact on human health.

The Cancer Genome Atlas (TCGA) has generated comprehensive molecular profiles including somatic mutation, copy number variation, gene expression, DNA methylation, microRNA expression, and protein expression for more than 30 different human tumor types. Those large datasets provided a great opportunity to examine the global landscape of aberrations at DNA, RNA, and protein levels.

In clinical areas the appearance of high-throughput sequencing technology called RNA-Seq has provided physicians with gene-expression data that allow more precise diagnosis and determination of the patient’s state from a molecular point of view. Since cancer is a heterogeneous disease driven by diverse genomic alterations, the analysis of gene-expression data obtained from tumor tissue samples allows the study of molecular factors contributing to disease progression over time or influencing patient’s survival. Gene expression data contain valuable information on the levels of differential activation of the genes involved in the development and evolution of cancer. If that information is extracted effectively, it can leverage precise diagnostic methods leading to greater efficacy of treatments and better prognosis.

TCGA RNA-Seq gene expression data helps discover features that can distinguish different tumor types. Those features may serve as biomarkers for tumor diagnosis and/or potential targets for drug development.

Not all cancer patients get diagnosed at the early stage. Our aim is to build a deep learning model based on TCGA RNA-Seq dataset to identify the types and subtypes of cancer and help medical service providers save their valuable time.

1. **Literature Survey:**

Discuss published information in a subject area

(OR)

Information in a particular subject area within a certain time period

(OR)

Summary of the sources,

It should have an organizational pattern and a combination of both summary and synthesis,

Summary is a recap of the important information of the source, and a-

-synthesis is a re-organization, or a reshuffling, of that information

(OR)

1. **Objectives:**

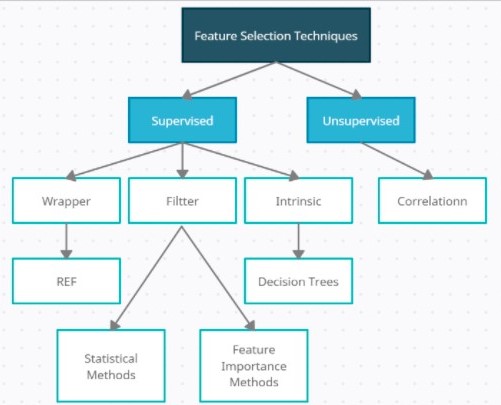
The early detection and diagnosis of a cancer is crucial for clinical treatment of patients.

Our aim is to build a deep learning model to identify cancer and its subtypes and help medical service providers save and utilize time spent in diagnosis for patient care and treatment.

1. **Methodology:**

a. Feature Selection:

**Feature selection** is the process of reducing the number of input variables when developing a predictive model. It is desirable to reduce the number of input variables to both reduce the computational cost of modeling and, in some cases, to improve the performance of the model. It can be supervised or unsupervised.



b. Feed Forward Neural Network (FFNN):

FFNN is an artificial neural network in which the connections between nodes does not form a cycle. It is the simplest form of neural network as information is only processed in one direction. While the data may pass through multiple hidden nodes, it always moves in one direction and never backwards.

FFNN is commonly seen in its simplest form as a single layer [perceptron](https://deepai.org/machine-learning-glossary-and-terms/perceptron). In this model, a series of inputs enter the layer and are multiplied by the weights. Each value is then added together to get a sum of the weighted input values. If the sum of the values is above a specific threshold, usually set at zero, the value produced is often 1, whereas if the sum falls below the threshold, the output value is -1. The single layer perceptron is an important model of feed forward neural networks and is often used in classification tasks.

In multi-layered perceptron, the process of updating weights is nearly analogous, however the process is defined more specifically as back-propagation. In such cases, each hidden layer within the network is adjusted according to the output values produced by the final layer.

Diagram, schematic

Description automatically generated

c. Generative Adversarial Networks (GAN):

GAN is a class of machine learning frameworks, it consists of two neural networks, discriminator and generator, these neural networks contest with each other in the form of a zero-sum game, where one agent’s gain is another agent’s loss.

The generative network generates candidates (images) while the discriminative network evaluates them.

Working of GAN:

The generator takes in a vector of random variables, Z, and produces an image, I.

The discriminator takes in the image, I, and produces a single output, p, deciding the probability if the image is real or fake.  
  
When:   
p=1, the discriminator believes that the image is real, and   
p=0, the discriminator believes that the image is fake

Diagram

Description automatically generated

1. **Modules identified:**

1. Study of the RNA Sequence Data Set

2. Data Pre-Processing

3. Feature Selection

4. Model building

5. Testing and Deployment

1. **Work progress and plan**

The current status of project and chart showing the monthly plan of project execution.

1. **Conclusions:**

The Generative Adversarial Network (GAN) based deep learning model built using TCGA RNA-Seq dataset will help identify cancer and its subtypes in a precise manner, thereby helping medical service providers save time.

1. **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] 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 multi-omics data. *Computational and Structural Biotechnology Journal*, *19*, 4404-4411.

[3] 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.

[4] López-García, G., Jerez, J. M., Franco, L., & Veredas, F. J. (2020). Transfer learning with convolutional neural networks for cancer survival prediction using gene-expression data. *PloS one*, *15*(3), e0230536.

[5] Chaudhari, P., Agrawal, H., & Kotecha, K. (2020). Data augmentation using MG-GAN for improved cancer classification on gene expression data.

*Soft Computing*, *24*(15), 11381-11391.

[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] Kourou, K., Rigas, G., Papaloukas, C., Mitsis, M., & Fotiadis, D. I. (2020). Cancer classification from time series microarray data through regulatory dynamic bayesian networks. *Computers in biology and medicine*, *116*, 103577

[8] Abdulqader, D. M., Abdulazeez, A. M., & Zeebaree, D. Q. (2020).

Machine learning supervised algorithms of gene selection: A review. *Machine Learning*, *62*(03).