VISVESVARAYA TECHNOLOGICAL UNIVERSITY

CONVOLUTON NEURAL NETWORK FOR CANCER SUBTYPE PREDICTION

PROJECT REPORT

Submitted by

AKHILESHKUMAR PATIL	1RF19IS005
K. SAI SUDHIR	1RF19IS020
NARESH SANTOSH SHET	1RF19IS030
RAKESH U	1RF19IS039

Under the Guidance of

Dr. Kirankumar Kataraki
Assistant Professor,
Information Science and Engineering, RVITM

in partial fulfillment for the award of degree

of

Bachelor of Engineering

in

Information Science and Engineering

RV INSTITUTE OF TECHNOLOGY AND MANAGEMENT,
BANGALORE-560076

2022-23

RV INSTITUTE OF TECHNOLOGY AND MANAGEMENT, BANGALORE - 560076
(Affiliated to VTU, Belgaum)

DEPARTMENT OF INFORMATION SCIENCE AND ENGINEERING

CERTIFICATE

Certified that the project work titled **'Convolution Neural Network for Cancer Subtype Prediction'** is carried out by **Akhileshkumar Patil (1RF19IS005), K. Sai Sudhir (1RF19IS020), Naresh Santosh Shet (1RF19IS030), Rakesh U(1RF19IS039),** who are bonafide students of RV Institute of Technology and Management, Bangalore, in partial fulfillment for the award of degree of **Bachelor of Engineering** in **Information Science and Engineering** of the Visvesvaraya Technological University, Belagavi during the year **2022-2023**. It is certified that all corrections/suggestions indicated for the internal Assessment have been incorporated in the report deposited in the departmental library. The project report has been approved as it satisfies the academic requirements in respect of project work prescribed by the institution for the said degree.

Signature of Guide **Signature of Head of the Department** **Signature of Principal**

External Viva

Name of Examiners	**Signature with date**
1	
2	

RV INSTITUTE OF TECHNOLOGY AND MANAGEMENT, BANGALORE -560059
(Affiliated to VTU, Belagavi)

DEPARTMENT OF INFORMATION SCIENCE AND ENGINEERING

DECLARATION

We, **K.Sai Sudhir,Rakesh U, Akhileshkumar Patil, Naresh Santosh Shet** the students of eighth semester B.E.,**Information Science and Engineering,** hereby declare that the project titled **"Convolution Neural Network for Cancer Subtype Prediction"** has been carried out by us and submitted in partial fulfillment for the award of degree of Bachelor of Engineering in **Information Science and Engineering.** We do declare that this work is not carried out by any other students for the award of degree in any other branch.

Place: Bangalore **Name** **Signature**

Date:

1. K. Sai Sudhir

2.Naresh Santosh Shet

3.Rakesh U

4.AkhileshKumar Patil

ACKNOWLEDGEMENT

The satisfaction and the euphoria that accompany the successful completion of any task would be incomplete without the mention of the people who made it possible. The constant guidance of these persons and encouragement crowned our efforts with success and glory. Although it is not possible to thank all the members who helped for the completion of the phase -2 of the project individually, we take this opportunity to express our gratitude to one and all.

We would like to thank the **VTU, Belagavi,** for having this project work as part of its curriculum, which gave us a wonderful opportunity to work on our research and presentation abilities.

We are indebted with a deep sense of gratitude for the constant inspiration, encouragement, timely guidance, and valid suggestions given to us by our guide **Dr. Kirankumar Kataraki,** Assistant Professor, Department of ISE, R V Institute of Technology and Management.

We wish to place on record our grateful thanks to **Dr. Latha C A,** Professor & HOD, Department of ISE, R V Institute of Technology and Management, for the constant encouragement provided to us.

We express our sincere gratitude to **Dr. Jayapal R,** Principal, R V Institute of Technology and Management for the support and encouragement.

We are thankful to the Project Coordinator and all the staff members of the department for providing relevant information and helping in different capacities in carrying out this phase - 2 project.

Last, but not least, we owe our debts to our parents, friends and those who directly or indirectly have helped us to make the phase - 2 project work a success.

AKHILESHKUMAR PATIL	1RF19IS005
K. SAI SUDHIR	1RF19IS020
NARESH SANTOSH SHET	1RF19IS030
RAKESH U	1RF19IS039

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 project 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 goal of cancer diagnosis is to classify tumors and identify indicators for each malignancy so that construct a learning system that can detect cancer early on.

In the field of cancer type prediction using gene expression data, various CNN (Convolutional Neural Network) models have been proposed. Each model is designed to tackle specific aspects related to modeling gene expression data and improving the accuracy of cancer type prediction

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 layers. Softmax is the activation function used for the last dense layer. To avoid overfitting a dropout rate of 0.15 is used. The model provides a test accuracy of 73.87%.

List of Publication

Akhileshkumar Patil, K.Sai Sudhir, Naresh Santosh Shet and Rakesh U "Convoluton neural network for Cancer subtype Prediction",Proceedings of National Conference on Degital Technology and Management, Bengaluru.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

List of symbols, Acronyms and Nomenclature

Symbols, Acronyms & Nomenclature	Definition	Page. No
TCGA RNA Seq Dataset	The Cancer Genome Atlas (TCGA) RNA (Ribonucleic acid)Seq Dataset, a landmark cancer genomics program, molecularly characterize 33 cancer types.	1
CNN	A Convolutional Neural Network, also known as CNN specializes in processing data that has a grid-like topology, such as an image.	1
MLP	A multilayer perceptron (MLP) is a fully connected class of artificial neural network (ANN). The term MLP is used to refer to networks composed of multiple layers of perceptrons.	2
KEGG	KEGG is a database resource for understanding high-level functions.	2
PAM50	PAM50 tests a sample of the tumor for a group of 50 genes.	5

PCR	PCR allows rapid detection of cancer DNA released in blood stream.	6
ATAM	This Method is a method used to evaluate the quality attributes(such as performance, availability, and security) of software architectures	9
NFR	Classifier uses information retrieval methods to find and identify NFRs.	10
Precision	Precision is a measure of how many of the positive predictions made are correct.	28
Recall	Recall is a measure of how many of the positive cases the classifier correctly predicted.	28
F1 Score	F1-Score is a measure combining both precision and recall.	28

Chapter 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% [1].

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 [2].

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 [3].

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[4].

1.1 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

1.2 Motivation

With today's technology, doctors can replace every part of the human body, from bones to organs, hands and face except brain and lungs. Hence early detection of damage in the lungs or brain should be recognized to improve the survival rate of human beings. This is the main motivation of this project. There are many techniques to diagnose lung cancer such as Chest Radiograph (X-Ray), Computed Tomography (CT), Magnetic Resonance Imaging (MRI) etc. But even after analyzing these reports, doctors may not accurately predict the stage of cancer or size of tumor. Therefore, there is a great need for a new technology i.e., Image Processing Techniques which is a good tool to improve manual analysis and to predict more accurately the size of tumor cells[5].

Biological pathway is an important curated knowledge of biological processes. Thus, cancer subtype classification based on pathways will be very useful to understand differences in biological mechanisms among cancer subtypes[3]. However, pathways include only a fraction of the entire gene set, only one-third of human genes in KEGG, and pathways are fragmented. For this reason, there are few computational methods to use pathways for cancer subtype classification[6].

1.3 Organization of Report

- **Chapter 1:**

This chapter covers the key aspects of gesture language recognition starting with a brief introduction to the objectives and motivations of the proposed system.

- **Chapter 2:**

In this chapter it contains the previously published works on gesture language. The purpose of this literature review is to gain an understanding of the existing research and debates relevant to the gesture recognition using deep learning, this chapter also contains the importance of the project.

- **Chapter 3:**

Includes the detailed examination of the system to understand its nature and fundamental or to determine its essential features.

- **Chapter 4:**

Represents the project in various designs, System architecture and Models of project.

- **Chapter 5:**

It includes the Dataset usage, Pre processing and Implemetation.

- **Chapter 6:**

In this chapter results been discussed.

- **Chapter 7:**

Concludes the project and provides the scope to which the project can be extended.

Chapter 2

Literature survey

The proposed system aims to accurately identify and predict the subtypes of various cancerous tumors using a deep learning-based Convolutional Neural Network (CNN)[4]. The system undergoes pre-processing steps such as handling missing data, feature selection using Recursive Feature Elimination, and normalization of selected genes. Heat maps are generated to visualize the gene expression data. The project leverages advancements in deep learning and computer vision to provide a valuable tool for cancer diagnosis and subtype prediction. By employing this system, individuals affected by cancer can gain insights into their specific cancer subtype, leading to improved medical care and personalized treatment strategies.

2.1 System study

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[10].

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 neighbor (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 colors 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 tumors from metastatic tumors. Finally, their deep learning model surpassed a machine learning ensemble approach on a variety of criteria's.

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 utilizing 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 remodeled 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 recognize twenty-one kinds of tumors 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 categorized 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. The author of paper [9] have proposed 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 [10] 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. The auther of paper [11] On several datasets of microarray, it has been discovered that identifying both FCR and PCR genes improves accuracy rate.

2.2 Proposed Work

The proposed work in a project refers to the planned activities and tasks that will be undertaken to achieve the project's goals and objectives. It encompasses various elements, including the project scope, which defines the boundaries and extent of the project. Additionally, the proposed work outlines specific objectives, which are the desired outcomes to be achieved within the project's timeline. A work breakdown structure (WBS) is used to break down the project into manageable tasks, phases, and deliverables. Milestones are established to mark significant points in the project's progress and track its completion. A detailed timeline and schedule are developed to ensure timely task completion and manage dependencies. Resources and budgetary considerations are identified to allocate the necessary personnel, equipment, and materials. The author of paper [12] proposed on Risk assessment is conducted to identify potential obstacles or challenges that may arise during project execution. Finally, the proposed work outlines the methodology or approach that will be followed, including the techniques, tools, or frameworks to be utilized. Altogether, the

proposed work serves as a comprehensive roadmap, providing a clear direction and plan for successfully executing the project and achieving its desired outcomes.

2.2.1 Problem Statement

The problem statement of this project is to develop and test an approach that can accurately identify and predict the subtypes of various cancerous tumors. The goal is to create a model that can analyze medical data, such as imaging or genetic data, and accurately classify the tumors into specific subtypes. The approach should aim to improve the accuracy and efficiency of tumor subtype identification, which can have significant implications for personalized treatment strategies and patient outcomes.

2.2.2 Existing System

In the field of cancer type prediction using gene expression data, various CNN (Convolutional Neural Network) models have been proposed. Each model is designed to tackle specific aspects related to modeling gene expression data and improving the accuracy of cancer type prediction.

One important aspect that has been addressed is the input gene order. The arrangement of genes in the input data can have an impact on the performance of the CNN models. Earlier methods have proposed techniques to optimize the arrangement of genes in a way that leads to the best prediction results.

One approach to address input gene order is by considering the chromosomal positions of the genes. Genes are located on chromosomes, and their positions on the chromosomes can provide valuable information about their interactions and functional relationships. Therefore, some methods propose ordering the genes based on their chromosomal positions.

The author of paper [13] proposed on genes based on chromosomal positions, the CNN models can potentially capture spatial dependencies and patterns that exist in the gene expression data. This ordering allows the models to effectively learn and extract relevant features from the data, improving the predictive performance.

Optimizing the gene arrangement based on chromosomal positions can be achieved through various techniques such as sorting the genes in ascending or descending order according to their positions, or using more advanced algorithms that take into account additional factors such as gene interactions or functional annotations.

By incorporating gene order optimization techniques into the CNN models, researchers aim to enhance the ability of the models to capture the underlying patterns and relationships in the gene expression data. This ultimately contributes to improving the accuracy and reliability of cancer type prediction.

2.2.3 Proposed System

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 Fig 2.1.

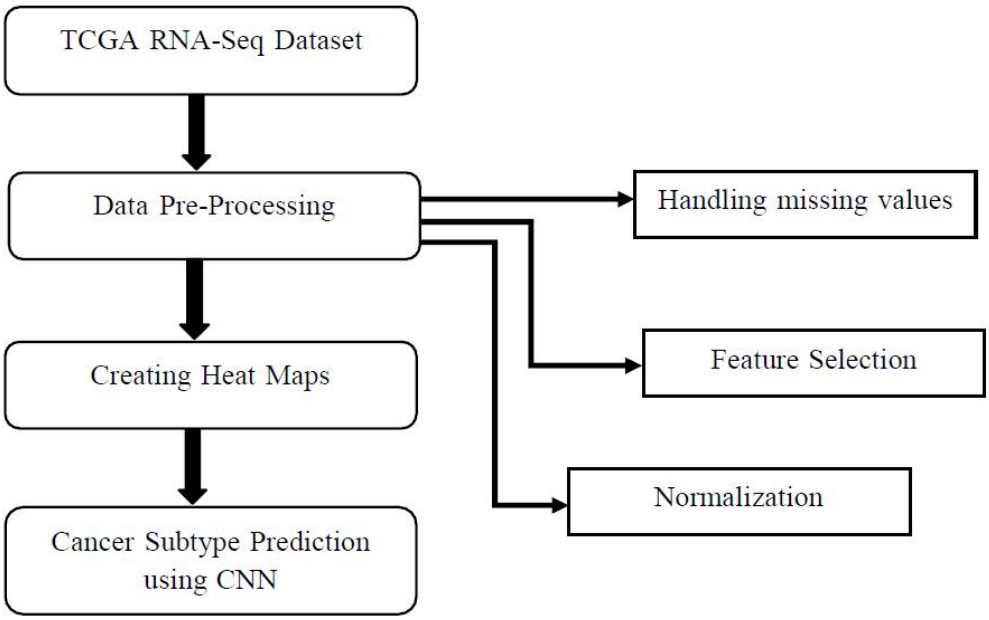


Fig 2.1 Flow of Data

I. Pre-processing:

The process shown in Fig 2.1 converting raw data into a comprehensible format is known as data pre-processing. Some of the pre-processing methods used are:

Missing Data:

Missing values refer to the absence of data for a variable in an observation. Missing data is a common occurrence in datasets and can have a significant impact on data analysis and the inferences drawn from it. Dealing with missing values is crucial to ensure the accuracy and reliability of data analysis results.

In data analysis, one approach to handling missing values is to drop the observations or variables that contain missing values. The author of paper [14] proposed on pandas library in Python provides the dropna() method, which allows for the removal of rows or columns containing null values from a Data Frame. Dropping missing values can be an effective strategy when the missing values are relatively few or randomly distributed. By removing observations or variables with missing values, you eliminate the potential bias that missing data can introduce into the analysis. However, dropping missing values may not always be the best approach, especially when missingness is not random or when a large portion of the data is missing. In such cases, dropping missing values may result in a significant loss of information and potential biases in the analysis. Alternative methods for handling missing data include imputation techniques, where missing values are replaced with estimated values based on patterns or relationships in the data. Imputation methods such as mean imputation, median imputation, or predictive modeling can help retain more observations and maintain the integrity of the dataset.

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.

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.

II. Heat Maps:

Heat maps are a widely used 2D information visualization technique that allows us to represent the intensity of an event or value using colors. In the context of analyzing gene

expression data, heat maps provide a visually appealing way to depict the expression levels of genes across different samples or patients.

To create a heat map, the first step is to preprocess the data present in a CSV file. Typically, this involves transposing the data so that the patient IDs are represented in rows and the different types of genes are represented in columns. This rearrangement ensures that the gene values of each patient are organized in a structured format.

Once the data is prepared, it is fed into a NumPy array, The author of paper [15] proposed onis a powerful numerical computing library in Python. NumPy arrays provide efficient storage and manipulation of multi-dimensional data.

The next step involves utilizing the Matplotlib library, specifically the function imshow(), to create the heat map image from the 2-dimensional NumPy array. Matplotlib is a versatile plotting library that offers a wide range of tools for visualizing data.

The imshow() function takes the NumPy array as input and generates an image where each cell of the array corresponds to a pixel in the image. The intensity of the color assigned to each pixel reflects the value of the corresponding gene expression level. Higher values are typically associated with brighter or warmer colors, while lower values are represented by darker or cooler colors.

The author of paper [16] proposed on Heat maps provide an intuitive and concise way to identify patterns and variations in gene expression across samples or patients. They allow researchers and analysts to quickly grasp the overall trends and differences in gene expression profiles, facilitating further exploration and analysis.

2.3 Scope of the project

The scope of this project is to develop a deep learning and computer vision-based model that can predict the subtype of cancer. The aim is to leverage advancements in technology to aid both common people and the medical field in identifying and understanding different subtypes of cancer.

Deep learning, a subset of machine learning, has shown great potential in various domains, including medical imaging analysis. The author of paper [17] proposed on utilizing deep learning algorithms and computer vision techniques, we can train a model to analyze medical images, such as histopathological images or radiological scans, and classify them into specific cancer subtypes.

The model will be trained on a large dataset of annotated cancer images, where each image is associated with a specific subtype. Through a process of feature extraction and learning, the model will learn to recognize patterns and characteristics specific to each subtype.

Once trained, the model can be deployed as a user-friendly application or tool. Users, including both medical professionals and individuals seeking information about their cancer subtype, can upload their medical images or input relevant data. The model will then process this input and provide a prediction of the cancer sub.

Chapter 3

Theory and fundamentals of area related to problem statement

Cancer subtypes refer to the distinct categories or classifications of cancer based on specific characteristics such as genetic alterations, molecular profiles, or clinical features. Understanding the theory and fundamentals related to cancer subtypes is crucial for accurate diagnosis, personalized treatment strategies, and improved patient outcomes.

At the core of cancer subtype classification is the understanding that cancer is not a single homogeneous disease but a complex group of diseases with unique molecular signatures and behaviors. Advances in molecular biology and genomics have enabled researchers to identify specific genetic mutations, gene expression patterns, and molecular alterations that differentiate cancer subtypes. The author of paper [18] proposed on discoveries have revolutionized our understanding of cancer biology and opened avenues for targeted therapies and precision medicine.

The study of cancer subtypes involves various fundamental concepts. One key aspect is the identification of biomarkers, which are specific molecules or genetic markers associated with particular cancer subtypes. The author of paper [19] proposed onBiomarkers can serve as diagnostic tools, helping clinicians determine the specific subtype of cancer a patient has. They can also guide treatment decisions by predicting response to certain therapies or prognosis.Additionally, cancer subtypes are often characterized by distinct molecular pathways and signaling networks that drive their growth and progression. Understanding these underlying molecular mechanisms is crucial for developing targeted therapies that can selectively inhibit or exploit these pathways, thereby improving treatment efficacy and reducing side effects.

The field of cancer subtype research also encompasses the concept of heterogeneity, which refers to the presence of diverse cell populations within a tumor. Tumors can consist of different subpopulations of cancer cells with varying genetic mutations or expression patterns, leading to differences in treatment response and disease progression. Recognizing and characterizing this intratumorally heterogeneity is essential for devising effective therapeutic strategies. The author of paper [20] proposed on discoveries have revolutionized our understanding of cancer biology and opened avenues for targeted therapies and precision medicine.

Furthermore, the concept of cancer subtypes extends beyond molecular and genetic features. Clinical characteristics such as tumor stage, histology, and patient demographics can also contribute to subtype classification. Integrating clinical and molecular information enables a more comprehensive understanding of cancer subtypes and facilitates personalized treatment decisions.

3.1 System Requirement Specification

System Requirement Specification is a structured collection of information that embodies the requirements of the system. It describes all the software's data, functional and behavioral requirements under production or development.

3.1.1 Hardware Requirements

- Processor: Intel Pentium CPU 2.6 GHz or AMD Athlon 64 (K8) 2.6 GHz
- 4 GB RAM
- Mouse
- Keyboard

3.1.2 Software Requirements

- Windows 10
- Python v3.4
- Open CV
- TensorFlow
- Keras

3.1.3 Functional Requirements

The functional requirements for a system describe what the system should do. These requirements depend on the type of software being developed, and the general approach taken by the organization when writing requirements. The functional system requirements describe the system function in detail, its inputs and outputs, exceptions, and so on.

Functional requirements are as follows:

- Accurate Dataset
- Access to the data

- Detected data analysis

3.1.4 Non-Functional Requirements

Non-functional requirements (NFRs) describe important constraints upon the development and behavior of a software system. They specify a broad range of qualities such as security, performance, availability, extensibility, and portability. These qualities play a critical role in architectural design and should therefore be considered and specified as early as possible during system analysis.

The non-functional requirements are as follows:

- **Elicitation techniques:** Elicitation methods are essential in the process of gathering requirements from stakeholders for the successful development of a system. These methods often employ various techniques such as creative brainstorming, the use of checklists, and Non-Functional Requirement (NFR) templates to prompt stakeholders to provide their input. One such approach is the Win-Win method, which involves the use of generic checklists and encourages stakeholders to contribute, prioritize, and negotiate NFRs that are deemed crucial for the system's success. Another approach is the Architectural Assessment Method (ATAM), which utilizes utility trees to model NFRs, allowing stakeholders to describe quality requirements within a hierarchical abstraction of high-level goals.

  The Win-Win method provides a structured framework for stakeholders to collaborate and identify important NFRs. It involves the use of checklists that cover a wide range of potential requirements, ensuring that all relevant aspects are considered. Stakeholders are encouraged to contribute their perspectives and insights, which leads to a comprehensive understanding of the system's non-functional needs. Through a process of prioritization and negotiation, stakeholders work together to establish a consensus on the most critical NFRs. This approach fosters a sense of shared ownership and aligns stakeholders' expectations, ultimately leading to a more successful system development.

  On the other hand, the ATAM approach employs utility trees to model NFRs. A utility tree is a graphical representation of quality requirements arranged in a hierarchical structure, where high-level goals are broken down into more specific attributes. Stakeholders participate in describing these attributes, allowing for a detailed and comprehensive representation of the system's quality needs. The utility tree

facilitates the analysis and evaluation of NFRs by providing a visual framework that captures the relationships and dependencies among different requirements. It enables stakeholders to prioritize and assess the impact of each NFR on the overall system architecture. This method enhances communication and understanding between stakeholders, as the utility tree serves as a common language for discussing and evaluating non-functional aspects.

- **Detection techniques:** Aspect-Oriented Programming (AOP) has gained significant popularity in software development due to its ability to address cross-cutting concerns that span multiple modules or components of a system. In AOP, aspects are used to encapsulate and modularize such concerns, making the codebase more maintainable and reducing code tangling and scattering. As the adoption of AOP increases, researchers have explored techniques to detect low-level aspects in design and code, as well as to identify "early aspects" from requirements specifications.

  Detecting low-level aspects in design and code involves identifying patterns and structures that exhibit cross-cutting behavior. Researchers have developed various static and dynamic analysis techniques to identify such aspects. Static analysis techniques analyze the codebase without executing it, looking for common code structures and dependencies that indicate cross-cutting concerns. The author of paper [21] proposed on analysis techniques, on the other hand, observe the behavior of the system during runtime and identify aspects based on runtime information.

  The identification of "early aspects" from requirements specifications aims to capture cross-cutting concerns at an early stage of software development, even before the coding phase. This is important as addressing concerns at the requirements level can lead to more efficient and effective system design. Researchers have proposed techniques such as aspect mining and aspect weaving to extract aspects from textual requirements documents. These techniques analyze the requirements specification to identify recurring themes, patterns, or keywords that indicate potential cross-cutting concerns. By detecting and addressing these concerns early in the development process, developers can avoid the proliferation of tangled and scattered code and improve the overall quality and maintainability of the system.

  The development of techniques for detecting low-level aspects in design and code, as well as identifying "early aspects" from requirements specifications, is driven by the

desire to enhance the modularity and maintainability of software systems. By effectively capturing and encapsulating cross-cutting concerns, AOP techniques enable developers to better manage complexity and improve code comprehensibility. Furthermore, by addressing concerns at an early stage, the overall development process can be streamlined, leading to more efficient and robust software systems.

- **The NFR-Classifier:** The NFR-Classifier is a solution designed to address the challenge of recognizing and retrieving Non-Functional Requirement (NFR) types. It achieves this by leveraging a training set to discover a set of weighted indicator terms for each NFR type. This approach, however, has the limitation of being restricted to recognizing and retrieving only the NFR types for which it has been trained. Despite this limitation, it is not overly restrictive as the system design process typically focuses on a smaller subset of common NFR types rather than the extensive list of over one hundred and fifty types documented.

  Furthermore, the concept of cancer subtypes extends beyond molecular and genetic features. Clinical characteristics such as tumor stage, histology, and patient demographics can also contribute to subtype classification. Integrating clinical and molecular information enables a more comprehensive understanding of cancer subtypes and facilitates personalized treatment decisions.

  The author of paper [22] proposed on NFR-Classifier offers several advantages over the standard keyword method. One key benefit is the ability to automatically mine indicator terms from existing pre-categorized requirement specifications. This allows organizations to customize the classifier according to their specific terminologies and policies. By leveraging their own standard terminologies, organizations can ensure that the NFR-Classifier aligns with their established practices, promoting consistency and facilitating easier adoption of the tool within the organization.

  Moreover, the NFR-Classifier's use of weighted indicator terms provides a more nuanced approach compared to the simplistic keyword matching technique. The weighting allows for a more accurate representation of the importance or relevance of specific terms within each NFR type. This enables the classifier to make more informed decisions when categorizing and retrieving NFRs based on their associated indicator terms. As a result, the NFR-Classifier can offer improved precision and recall in identifying NFRs, enhancing the effectiveness of the system design process.

The Win-Win method provides a structured framework for stakeholders to collaborate and identify important NFRs. It involves the use of checklists that cover a wide range of potential requirements, ensuring that all relevant aspects are considered. Stakeholders are encouraged to contribute their perspectives and insights, which leads to a comprehensive understanding of the system's non-functional needs. Through a process of prioritization and negotiation, stakeholders work together to establish a consensus on the most critical NFRs. This approach fosters a sense of shared ownership and aligns stakeholders' expectations, ultimately leading to a more successful system development.

Chapter 4

Design

The system architecture diagram for this project is shown in the Fig below. Data is split into training and testing. The dataset is tried on various supervised algorithms and accuracy is calculated. Based on which algorithm gives the best accuracy the model is built. Then the model is tested on the testing dataset and the results are verified. Finally, the model is run on real-time data and it predicts cancer as well as recommends therapy.

Fig 4.1 shows the architectural diagram of our ML system. It consists of different sub-modules which include splitting data into training and testing. The training data is used to build the model. Training data is given to the algorithm, the algorithm is further evaluated in the evaluation step. Now in the evaluation step, all the evaluation is done and further flow goes to the cancer report module which detects.

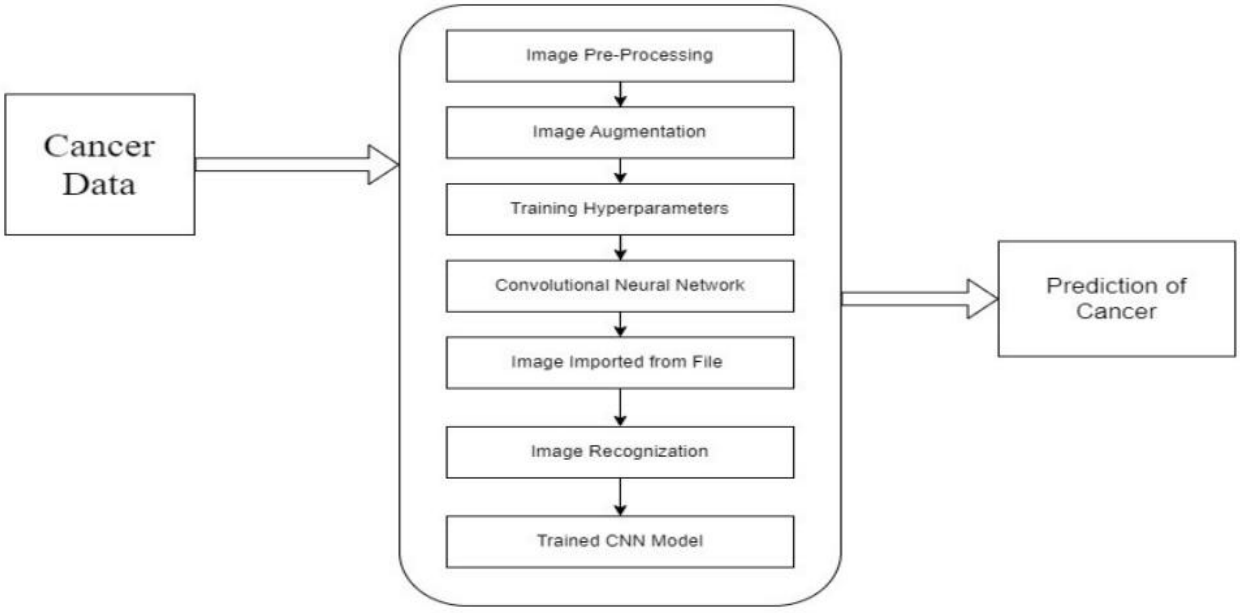


Fig 4.1 Design Overview

4.1 System Architecture

The architecture of a system reflects how the system is used and how it interacts with other systems and the outside world. It describes the interconnection of all the system's components and the data link between them. The architecture of a system reflects the way it is thought about in terms of its structure, functions, and relationships. In architecture, the term "system" usually refers to the architecture of the software itself, rather than the physical structure of the buildings or machinery. The architecture of a system reflects the way it is used and therefore changes as the system is used. For example, an airport may be designed using an architecture where the control tower and departures lounge are close together in the same building, while the control tower is further away in the same airport.

The system architecture diagram for this project is as shown in the Fig below. Data is split into training and testing. The dataset is tried on various supervised algorithms and accuracy is calculated. Based on which algorithm gives the best accuracy the model is built. Then the model is tested on the testing dataset and the results are verified. Finally, the model is run on real-time data and it predicts cancer as well as recommends therapy.

Fig 4.2 shows the architectural diagram of our ML system. It consists of different sub-modules which include splitting data into training and testing. The training data is used to build the model. Training data is given to the algorithm, the algorithm is further evaluated in the evaluation step. Now in the evaluation step, all the evaluation is done and further flow goes to the cancer report module which detects cancer and recommends treatment.

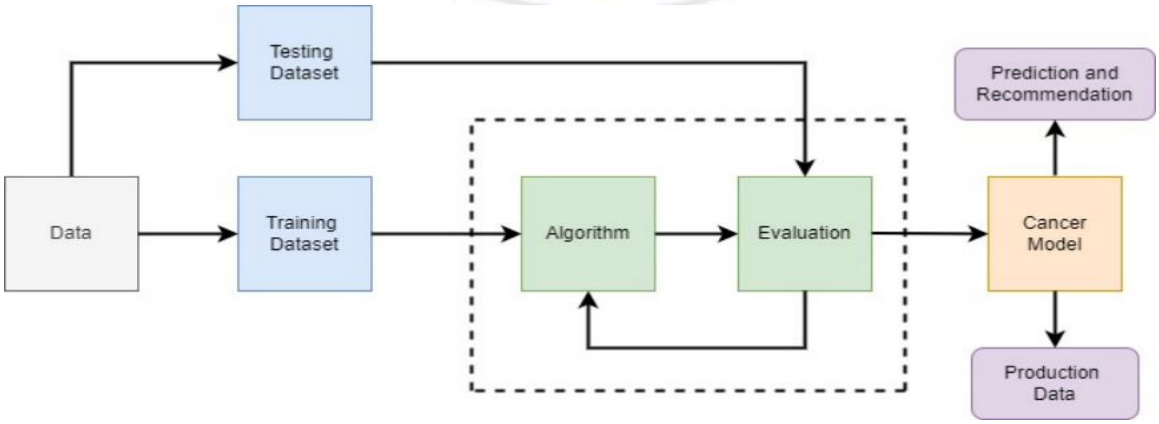


Fig 4.2 System Architecture

4.2 Modules of the Project

A structure chart is used to represent the control flows happening between the various modules of the system, this chart can be used to explain the various modules and their interactions with the other modules in the system. It can be utilized to explain what are the inputs given to each of the modules and what output are expected by them. In the structure, chart modules and functions are represented as squares and the lines between these modules represent the connection between them. The structure chart in Fig depicts how the processes are carried out in the system.

Fig 4.2 shows the complete data cleaning and preprocessing process. Data preprocessing has three major steps, understanding the dataset by exploratory data analysis, and splitting the dataset for training and testing. It is usually done in a ratio of 85:15. Finally the resolve is where actual data cleaning, encoding, and normalization happen. Here missing values are resolved, words are converted to numbers, dealing with outliners and lastly, useless values are eliminated.

1. **Data Pre-Processing**
2. **RNA-Sequencing**
3. **Feature Selection**
4. **Deployment**

4.3.1 Data Pre-Processing

Data pre-processing is a step in the data mining and data analysis process that takes raw data and transforms it into a format that can be understood and analyzed by computers and machine learning. Raw, real-world data in the form of text, images, video, etc., is messy. Not only may it contain errors and inconsistencies, but it is often incomplete and doesn't have a regular, uniform design.

Machines like to process nice and tidy information – they read data as 1s and 0s. so calculating structured data, like whole numbers and percentages is easy. However, unstructured data, in the form of text and images must first be cleaned and formatted before analysis.

Let's take a look at the established steps you'll need to go through to make sure your data is successfully preprocessed.

1. Data Integration
2. Data cleaning
3. Data transformation
4. Data reduction

Data Integration

Data integration is a crucial process in bringing together data from various disparate sources to create a unified view for users. Its primary objective is to make data more accessible, manageable, and meaningful to both systems and users. By integrating data, organizations can achieve numerous benefits, including cost reduction, resource optimization, improved data quality, and increased opportunities for innovation. The beauty of data integration lies in its ability to achieve these advantages without requiring significant changes to existing applications or data structures.

Effective data integration can lead to reduced IT costs. By consolidating data from different sources into a single data store, organizations can avoid the need for multiple data management systems and the associated costs of maintaining and managing them separately. Instead, a unified view enables streamlined data access, storage, and processing, resulting in cost savings and operational efficiencies.

Data integration also frees up resources within the organization. Without integration, data from different sources often requires manual effort and time-consuming processes to extract, transform, and load (ETL) it into a usable format. However, with proper integration, these tasks can be automated, allowing IT personnel to focus on more value-added activities such as data analysis, insights generation, and decision-making.

Improved data quality is another significant advantage of data integration. By integrating data from multiple sources, organizations can identify and resolve inconsistencies, redundancies, and inaccuracies. Data cleansing and standardization can be performed during the integration process, leading to a unified and high-quality data repository. Clean and reliable data ensures that subsequent analyses, reporting, and decision-making processes are based on accurate and consistent information.

Data integration also fosters innovation within organizations. By creating a unified view of data, it becomes easier to discover patterns, relationships, and insights that may not have been apparent when the data was siloed. Integrated data enables comprehensive analysis and facilitates data-driven decision-making, ultimately driving innovation and uncovering new opportunities for business growth.

Effective data integration can lead to reduced IT costs. By consolidating data from different sources into a single data store, organizations can avoid the need for multiple data management systems and the associated costs of maintaining and managing them separately. Instead, a unified view enables streamlined data access, storage, and processing, resulting in cost savings and operational efficiencies. . Its primary objective is to make data more accessible, manageable, and meaningful to both systems and users. By integrating data, organizations can achieve numerous benefits, including cost reduction, resource optimization, improved data quality, and increased opportunities for innovation. The beauty of data integration lies in its ability to achieve these advantages without requiring significant changes to existing applications or data structures.

Data Cleaning

Data cleaning, also known as data cleansing or data scrubbing, is a crucial step in the data preprocessing phase. It involves identifying and correcting or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. Data cleaning is essential to ensure the accuracy, reliability, and usability of the data for subsequent analyses and decision-making.

When combining multiple data sources, the likelihood of encountering duplicated or mislabeled data increases. These inconsistencies can lead to unreliable outcomes and inaccurate algorithmic results, even if they appear to be correct at first glance. Data cleaning aims to address these issues and improve the overall quality of the dataset.

The specific steps and techniques involved in the data cleaning process can vary depending on the characteristics and peculiarities of the dataset. There is no one-size-fits-all approach to data cleaning, as each dataset requires tailored treatment. However, it is crucial to establish a template or framework for the data cleaning process to ensure consistency and accuracy across different datasets.

Establishing a data cleaning process involves defining a set of steps and procedures to follow systematically. This may include identifying and handling missing values, removing

duplicate records, standardizing data formats, correcting inconsistencies or errors, validating data against predefined rules or constraints, and transforming data into a suitable format for analysis.

Data cleaning often requires using specific techniques or algorithms tailored to the type of data being cleaned. For example, in textual data, techniques such as spell-checking, stemming, or removing stop words may be applied. In numerical data, outliers or extreme values may be identified and treated appropriately. Domain-specific knowledge and expertise are often necessary to determine the appropriate cleaning methods and ensure the integrity and relevance of the data.

By conducting a data quality assessment, inconsistencies and issues within the dataset are identified. Data cleaning aims to rectify these problems, ensuring that the dataset is accurate, complete, and reliable. The outcome of the data cleaning process is a cleansed dataset that is ready for further analysis, modelling, or visualization.

Data Transformation

Data transformation is a fundamental process in preparing data for analysis and utilization. It involves converting raw data into a format that is suitable for machine processing and analysis. While feature transformation focuses on replacing attributes with mathematical functions, data transformation primarily focuses on making numerical and categorical data machineready.

The data transformation process is often referred to as extract/transform/load (ETL). It starts with the extraction phase, where data is identified and pulled from various source systems. The extracted data is then moved to a centralized repository or data storage. In the transformation phase, the raw data undergoes cleansing to address any inconsistencies, errors, or missing values. It is then transformed into a target format that can be used by operational systems, data warehouses, data lakes, or other repositories for business intelligence and analytics applications.

Data transformation encompasses several tasks, including converting data types, standardizing formats, aggregating or disaggregating data, normalizing values, and enriching the data with additional attributes. These transformations ensure that the data is consistent, accurate, and suitable for the intended use. For example, data types may need to be converted from text to numeric for mathematical calculations, or categorical variables may need to be one-hot encoded for machine learning algorithms.

Data transformation is crucial in various data-related processes such as data integration, data cleaning, data migration, data warehousing, and data wrangling. It plays a vital role in harmonizing data from disparate sources, resolving inconsistencies, and preparing data for analysis and reporting. Without proper transformation, data may be difficult to interpret, analyze, and derive insights from.

In the era of big data, where data volumes are growing rapidly, organizations need efficient methods to harness and leverage their data effectively. Data transformation is a critical component in this process as it ensures that data is accessible, consistent, secure, and trusted by business users. Proper data transformation enables organizations to generate timely and reliable insights, make informed decisions, and drive business success.

Data Reduction

Data reduction techniques play a crucial role in managing and analyzing large datasets. These techniques aim to reduce the volume of the original data while maintaining its integrity. By obtaining a reduced representation of the dataset, data reduction improves the efficiency of the data mining process without significantly impacting the analytical results. The primary objective of data reduction is to represent the dataset in a much smaller volume without compromising the essential information contained within it. Reducing the data size has several benefits, including improved computational efficiency and reduced storage requirements. Smaller datasets are more amenable to applying complex algorithms and analyses, which may be computationally expensive when dealing with large volumes of data. By conducting a data quality assessment, inconsistencies and issues within the dataset are identified. Data cleaning aims to rectify these problems, ensuring that the dataset is accurate, complete, and reliable. The outcome of the data cleaning process is a cleansed dataset that is ready for further analysis, modelling, or visualization.

4.3.2 RNA-Sequencing

RNA sequencing has significantly progressed, becoming a paramount approach for transcriptome profiling. The revolution from bulk RNA sequencing to single-molecular, single-cell, and spatial transcriptome approaches has enabled increasingly accurate, individual-cell resolution incorporated with spatial information. Cancer, a major malignant and heterogeneous lethal disease, remains an enormous challenge in medical research and clinical treatment.

As a vital tool, RNA sequencing has been utilized in many aspects of cancer research and therapy, including biomarker discovery and characterization of cancer heterogeneity and evolution, drug resistance, cancer immune microenvironment and immunotherapy, cancer neoantigens and so on. In this review, the latest studies on RNA sequencing technology and its applications in cancer are summarized, and future challenges and opportunities for RNA sequencing technology in cancer applications are discussed.

Data reduction can be achieved by reducing the number of rows (records) or columns (dimensions) in the dataset. Reducing the number of rows can involve techniques such as sampling or clustering to select a representative subset of the data. This subset retains the essential patterns and characteristics of the original dataset while reducing its size. Similarly, reducing the number of columns can involve feature selection or feature extraction techniques to identify the most relevant and informative attributes for analysis.One of the advantages of data reduction is that it does not significantly impact the results obtained from data mining. The analytical outcomes before and after data reduction are generally similar or almost the same. This is because data reduction techniques are designed to preserve the key patterns and relationships within the data, ensuring that the important information is retained while discarding redundant or irrelevant details.

In addition to improving computational efficiency, data reduction also simplifies the analysis process and enhances accuracy. When working with large volumes of data, especially in text analysis, much of the information may be superfluous or irrelevant to the research goals. Data reduction allows researchers to focus on the most relevant and meaningful aspects of the data, streamlining the analysis and reducing the potential for noise or irrelevant factors to influence the results.

Moreover, data reduction helps in cutting down on data storage requirements. Storing and managing large datasets can be resource-intensive and costly. By reducing the data size through techniques such as compression or summarization, organizations can significantly reduce the storage needs and associated costs while still preserving the essential information for analysis.

4.3.3 Feature Selection

A feature is an attribute that has an impact on a problem or is useful for the problem, and choosing the important features for the model is known as feature selection. Each machine learning process depends on feature engineering, which mainly contains two processes,

which are Feature Selection and Feature Extraction. Although feature selection and extraction processes may have the same objective, both are completely different from each other. The main difference between them is that feature selection is about selecting the subset of the original feature set, whereas feature extraction creates new features. Feature selection is a way of reducing the input variable for the model by using only relevant data to reduce overfitting in the model.

Recursive Feature Elimination

Recursive Feature Elimination (RFE) is a feature selection technique used to select a subset of features from a larger set of available features. It is a recursive and greedy optimization approach that iteratively selects features based on their importance or contribution to the model's performance.The RFE process starts by training an estimator, such as a machine learning algorithm, on the entire set of features. The importance or contribution of each feature is then determined using a coefficient attribute (e.g., coef_) or a feature importances attribute (e.g., feature_importances_). These attributes provide a measure of how much each feature influences the prediction or performance of the model.Next, the least important features are removed from the feature set. The estimator is retrained using the reduced feature set, and the feature importance is re-evaluated. This recursive process continues until a predetermined number of features or a desired level of feature importance is reached.During each iteration, RFE evaluates the performance of the estimator using the reduced feature set. The goal is to identify the subset of features that maximizes the performance of the model while minimizing the number of features. By recursively eliminating the least important features, RFE aims to find the optimal subset that retains the most relevant and informative features for the task at hand.

Forward selection

Forward selection is a feature selection technique that iteratively builds a model by adding one feature at a time based on its contribution to the model's performance. The process starts with an empty set of features and gradually selects the most relevant features to improve the model's performance. In each iteration of forward selection, the algorithm evaluates the performance of the model by adding one additional feature from the pool of remaining features. The feature that provides the most improvement in the model's performance, as determined by a predefined evaluation metric (e.g., accuracy, precision, or mean squared error), is selected and added to the feature set. This iterative process continues

until the addition of a new feature no longer improves the model's performance significantly. At this point, the algorithm stops, and the selected features form the final subset used for modeling.

4.3.4 Deployment

A deployment diagram in Fig 4.3 shows the configuration of run-time processing nodes and the components that live on them. Deployment diagrams are a kind of structure diagram used in modeling the physical aspects of an object-oriented system. They are often used to model the static deployment view of a system.

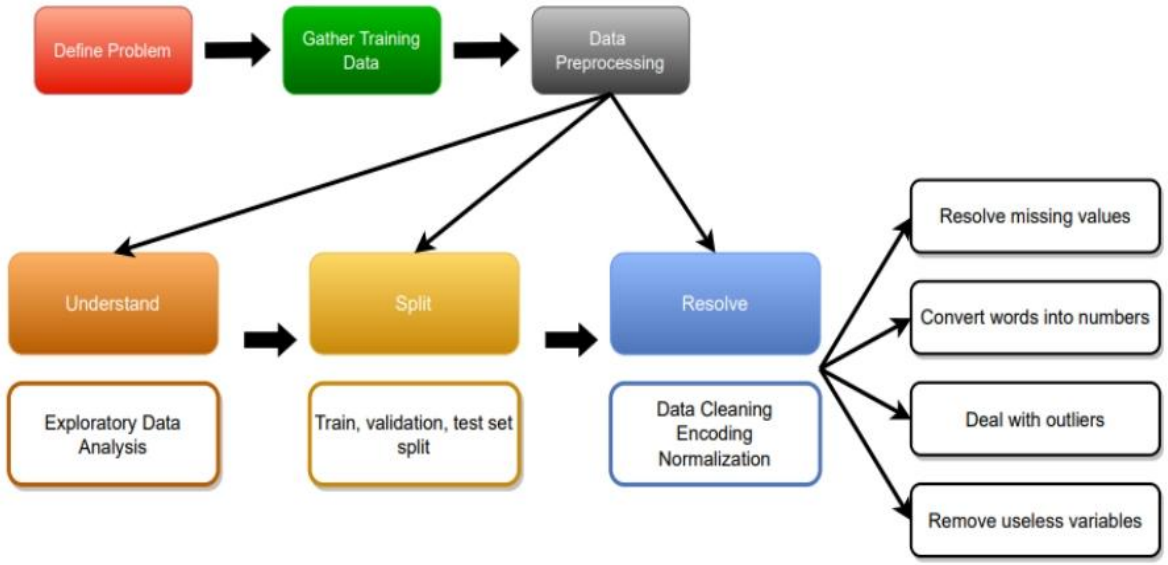


Fig 4.3 Deployment of Module

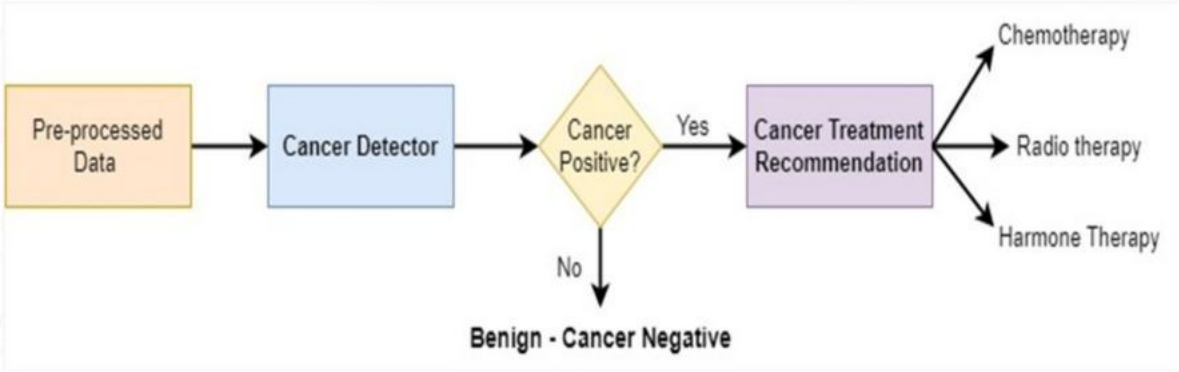


Fig 4.4 Flow of project

The Fig 4.4 above shows the major methods of the project. Every step is crucial and kept separately for simplicity of design and development. Only if cancer detector returns cancer positive then the data is further passed to cancer treatment recommendation module. If

not, the cancer detector returns cancer negative and terminates the process. Cancer treatment recommender suggests which is best for the input given.

4.4 Design Diagrams

4.4.1 Sequence Diagram

A sequence diagram is an interaction diagram. From the name, it is clear that the diagram deals with some sequences, which are the sequence of messages flowing from one object to another. Sequence Diagrams in Fig 4.5 captures the interaction that takes place in a collaboration that either realizes a use case or an operation (instance diagrams or generic diagrams) high-level interactions between user of the system and the system, between the system and other systems, or between subsystems (sometimes known as system sequence diagrams)

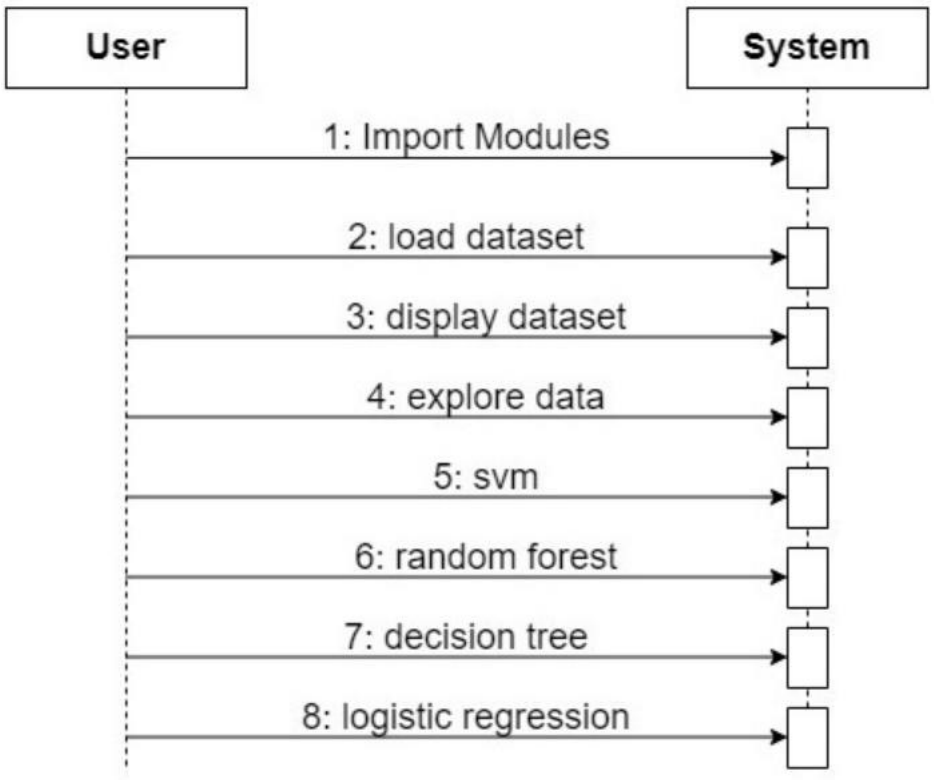


Fig 4.5 Sequence Diagram

4.4.2 Collaboration Diagram

Collaboration diagram is another form of interaction diagram. It represents the structural organization of a system and the messages sent/received. Structural organization consists of objects and links. The purpose of collaboration diagram is similar to sequence diagram. However, the specific purpose of collaboration diagram is to visualize the organization of objects and their interaction.

In Fig 4.6 collaboration diagram is used to show the relationship between the objects in a system. Both the sequence and the collaboration diagrams represent the same information but differently. Instead of showing the flow of messages, it depicts the architecture of the object residing in the system as it is based on object-oriented programming. An object consists of several features. Multiple objects present in the system are connected to each other. The collaboration diagram, which is also known as a communication diagram, is used to portray the object's architecture in the system.

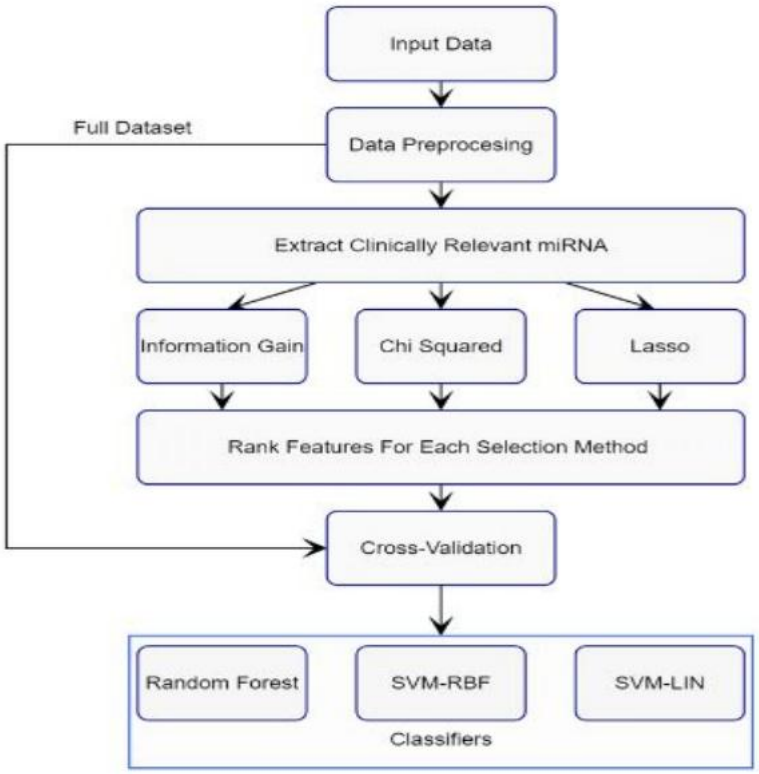


Fig 4.6 Collaboration Diagram

4.4.3 Data Flow Diagram

Workflow diagrams, specifically Data Flow Diagrams (DFDs), are graphical representations used to illustrate the flow of data within an information system. They provide a visual depiction of how data moves through various processes and components within a system, from input to output.

A DFD consists of different symbols such as rectangles, circles, and arrows, along with labels, to represent the data inputs, outputs, storage points, and the paths connecting them. The symbols and connections in a DFD demonstrate how data is processed, transformed, stored, or transmitted within the system.

DFDs can range from simple, high-level overviews of a process to more complex and detailed representations that delve into the intricacies of data handling. They are versatile and can be used to analyze existing systems or model new ones. DFDs are effective for both technical and non-technical audiences, as they provide a visual representation that can convey complex information more easily than textual descriptions.

One of the advantages of DFDs is their ability to communicate information that might be difficult to express in words alone. By visually mapping out the flow of data, DFDs enable stakeholders to understand how inputs are transformed into outputs and identify potential bottlenecks or areas for improvement within a system.DFDs are particularly useful for data flow-oriented software and systems, where understanding the movement of data is crucial. They can aid in identifying data dependencies, highlighting interactions between processes, and ensuring the smooth flow of information.However, it's important to note that DFDs may have limitations in certain contexts.

They are less applicable when visualizing interactive or real-time systems that involve dynamic interactions or complex user interfaces. Additionally, for database-oriented systems, other diagramming techniques such as Entity-Relationship Diagrams (ERDs) may be more appropriate for capturing the relationships between data entities.

Workflow diagram for detection of cancer

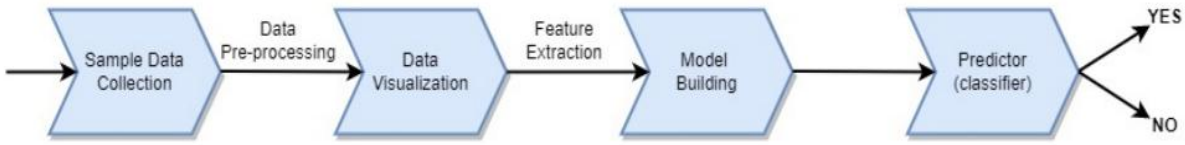


Fig 4.7 Cancer detection workflow

As shown in the Fig 4.7 the collected data undergoes several steps before the final prediction is made. Firstly, the collected data is fed into the system as input.

After the data is collected, it goes through a pre-processing step. Data pre-processing involves cleaning the data by handling missing values, removing outliers, or normalizing the data to ensure consistency and reliability. This step aims to prepare the data for further analysis and modeling.

Once the data is pre-processed, it is then visualized. Data visualization techniques are employed to represent the data in a graphical or visual format. Visualization helps in gaining insights and understanding the patterns, trends, and relationships within the dataset. It allows for a better comprehension of the data and aids in making informed decisions, such as selecting the most suitable algorithm for the given dataset.

Based on the extracted features from the pre-processed and visualized data, a model is built. This involves selecting an appropriate machine learning algorithm that is capable of learning patterns and relationships from the data. The model is trained using the labeled data, where the outcomes (cancer positive or negative) are already known.

Once the model is trained, it becomes the predictor. It takes in new, unseen instances of data and predicts whether the given case is cancer positive (YES) or cancer negative (NO) based on the learned patterns and relationships from the training data. The predictor uses the extracted features from the input data to make its predictions.

Chapter 5

Implementation

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 Fig 5.1

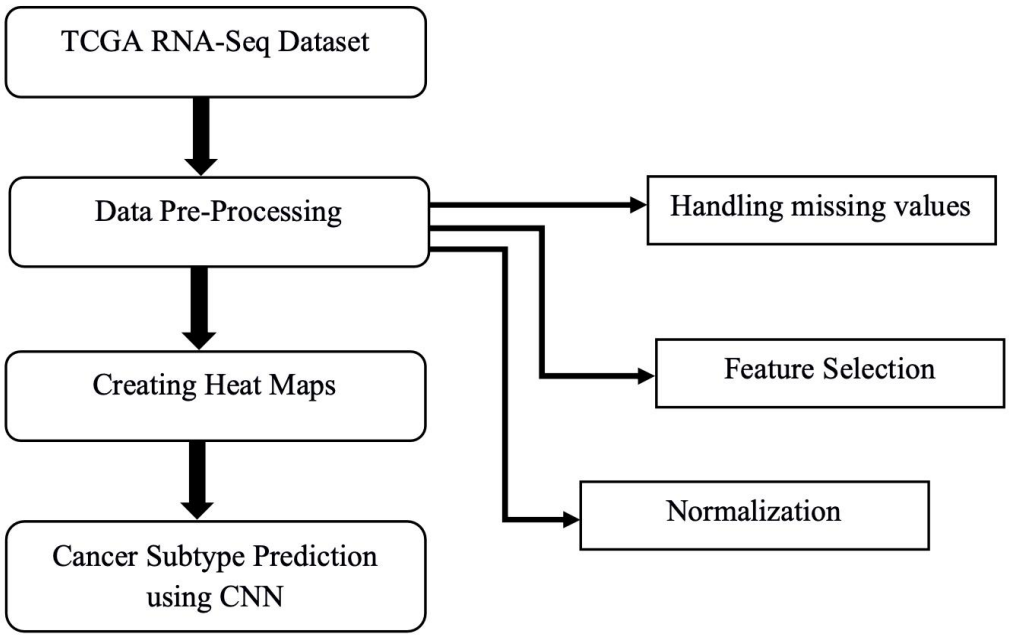


Fig 5.1 Diagram of proposed system

5.1 Dataset

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

Table 5.1 Screenshot of a part of 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

5.2 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:

Missing Data

Missing values refer to the absence of data for a variable in an observation. Missing data is a common occurrence in datasets and can have a significant impact on data analysis and the inferences drawn from it. Dealing with missing values is crucial to ensure the accuracy and reliability of data analysis results.

In data analysis, one approach to handling missing values is to drop the observations or variables that contain missing values. The pandas library in Python provides the dropna() method, which allows for the removal of rows or columns containing null values from a DataFrame. Dropping missing values can be an effective strategy when the missing values are relatively few or randomly distributed. By removing observations or variables with missing values, you eliminate the potential bias that missing data can introduce into the analysis. However, dropping missing values may not always be the best approach, especially when missingness is not random or when a large portion of the data is missing. In such cases, dropping missing values may result in a significant loss of information and potential biases in the analysis. Alternative methods for handling missing data include imputation techniques, where missing values are replaced with estimated values based on patterns or relationships in the data. Imputation methods such as mean imputation, median imputation, or predictive modeling can help retain more observations and maintain the integrity of the dataset.

Feature Selection

One popular feature selection method is Recursive Feature Elimination (RFE). RFE is an attribute selection approach that iteratively eliminates the lowest-ranked features until a desired number or a predefined set of attributes is achieved. The selection process starts with a model trained on all the available features. The importance or relevance of each feature is assessed using a ranking criterion, such as coefficients, feature importance, or information gain. The least important feature(s) are then removed, and the model is retrained on the reduced feature set. This process is repeated until the desired number of features is reached.

The Recursive Feature Elimination technique is applied to the TCGA RNA-Seq dataset, which likely contains a large number of genes or features (20,531 in this case). By applying RFE, the algorithm systematically eliminates the least important genes based on their relevance to the target variable or the model's performance. The process continues until the desired number of features, in this case, 50 genes, is selected.

The advantage of using RFE is that it considers the interaction and dependence between features during the elimination process. By iteratively eliminating the least important features, RFE aims to retain the most relevant and informative features for the model. This can help improve the model's performance, reduce overfitting, and enhance interpretability.

Normalization

Normalization is the process of converting data to a common scale. In this project, the 50 selected genes are normalized to a range of 0 to 255, ensuring that they have the same scale for consistency and comparability in further analysis or modeling.

5.3 HeatMaps

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.

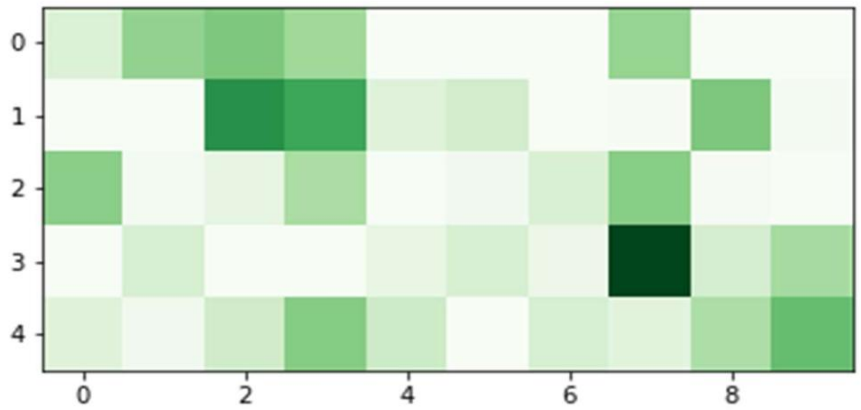


Fig 5.2 Heat Map of cancer type ACC

5.4 Model Architecture

The CNN architecture represented by Fig 5.1 and Fig 5.2 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 layers. Softmax is the activation function used for the last dense layer. To avoid overfitting, the dropout rate of 0.15 is used.

Table 5.2 CNN Model Architecture

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

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

5.5 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.

5.6 Testing

To ensure the accuracy and effectiveness of the articles written during the internship, I conducted a series of tests to validate the methods and options described in the articles. The following describes each test in more detail.

Unit Testing

For the unit tests, each method and option were individually tested to verify that it produced the expected results. This involved designing test cases that covered different scenarios and input variations. The tests assessed the accuracy, robustness, and efficiency of the methods and options. By comparing the actual output with the expected output, any discrepancies or issues in the implementation could be identified and addressed.

Integration testing was conducted to evaluate the interactions between different methods and options. This ensured that when combined, they seamlessly worked together and produced the desired outcome. The integration tests focused on verifying the compatibility and compatibility of the methods and options, identifying any potential conflicts or dependencies. By testing the system, it was possible to identify any issues that might arise due to the interaction between components. The test cases for integration testing were designed to cover different scenarios and workflows that involved multiple methods and options. The inputs and outputs of each component were carefully monitored to ensure that the integration was successful and that the desired results were obtained. . The tests assessed the accuracy, robustness, and efficiency of the methods and options. By comparing the actual output with the expected output, any discrepancies or issues in the implementation could be identified and addressed.

The unit tests and integration testing were essential for validating the functionality and reliability of the methods and options. They provided confidence in the accuracy and effectiveness of the implemented techniques. Any bugs, errors, or inconsistencies were detected and resolved, leading to improved performance, and ensuring that the methods and options could be relied upon for further analysis and decision-making.

Overall, the unit tests and integration testing played a crucial role in validating the methods and options described in the articles. By thoroughly testing the functionality, compatibility, and performance, it was possible to ensure the quality and reliability of the methods and options, providing a solid foundation for their application in practical settings.

Performance Testing

Performance testing involved subjecting the methods and optionsto different workloads, simulating real-world usage scenarios. This included testing the system under normal operating conditions as well as high load scenarios to evaluate its responsiveness and stability. The performance metrics measured included response time, throughput, resource utilization, and scalability. Various tools and techniques were employed to conduct the performance testing. Load testing tools were utilized to generate simulated user traffic and stress the system with high loads. This helped identify the system's breaking point and measure its ability to handle concurrent user requests.

During the performance tests, performance counters and monitoring tools were utilized to collect data on system performance, such as CPU and memory usage, network latency, and disk I/O. These metrics provided insights into the system's performance bottlenecks and resource utilization patterns.The results of the performance testing were analyzed to identify any performance issues or areas that required optimization. This could involve optimizing algorithms, improving database query performance, or optimizing resource allocation. By addressing the identified issues, the methods and options could be fine-tuned for improved speed and efficiency.

The performance testing also provided valuable insights into the system's scalability. By gradually increasing the load, the scalability of the methods and options could be assessed to ensure they could handle growing user demands without significant degradation in performance.

User Acceptance Testing

User acceptance testing involved providing the articles to the selected group of Python developers and asking them to review and provide feedback. The participants were instructed to assess the articles based on their understanding of the content, relevance to their needs, and overall usability.

During the testing, participants were encouraged to actively engage with the articles, ask questions, and provide their opinions on the presented information. Their feedback was collected through surveys, interviews, or direct discussions. The focus was on gathering qualitative feedback on aspects such as clarity of explanations, comprehensibility of examples, and relevance to their specific needs. The feedback gathered during user acceptance testing was valuable in several ways. Firstly, it provided insights into the

effectiveness of the articles in conveying the intended information to the target audience. It helped identify areas where the content might be confusing, lacking detail, or requiring further clarification.

Secondly, user acceptance testing allowed for the identification of any inaccuracies or inconsistencies in the articles. Developers were able to point out any technical errors or misleading information, enabling necessary corrections and improvements to be made. Additionally, user acceptance testing helped assess the overall usefulness of the articles. Participants could provide input on whether the content met their expectations, addressed their concerns, and provided practical insights or guidance for their Python development needs.

The feedback gathered from user acceptance testing was carefully analyzed and considered in the refinement and improvement of the articles. It helped ensure that the articles met the expectations of the target audience, provided accurate and reliable information, and effectively served their purpose in delivering valuable knowledge to Python developers.

Chapter 6

Results, discussions and inference

Snapshot 1: Model Accuracy

Accuracy snapshot 1 Fig 6.1 is the one metric for evaluating classification models. Informally, accuracy is the fraction of predictions our model got right. Accuracy comes out to 0.76, or 76% (76 correct predictions out of 100 total examples). That means our tumor classifier is doing a great job of identifying malignancies.

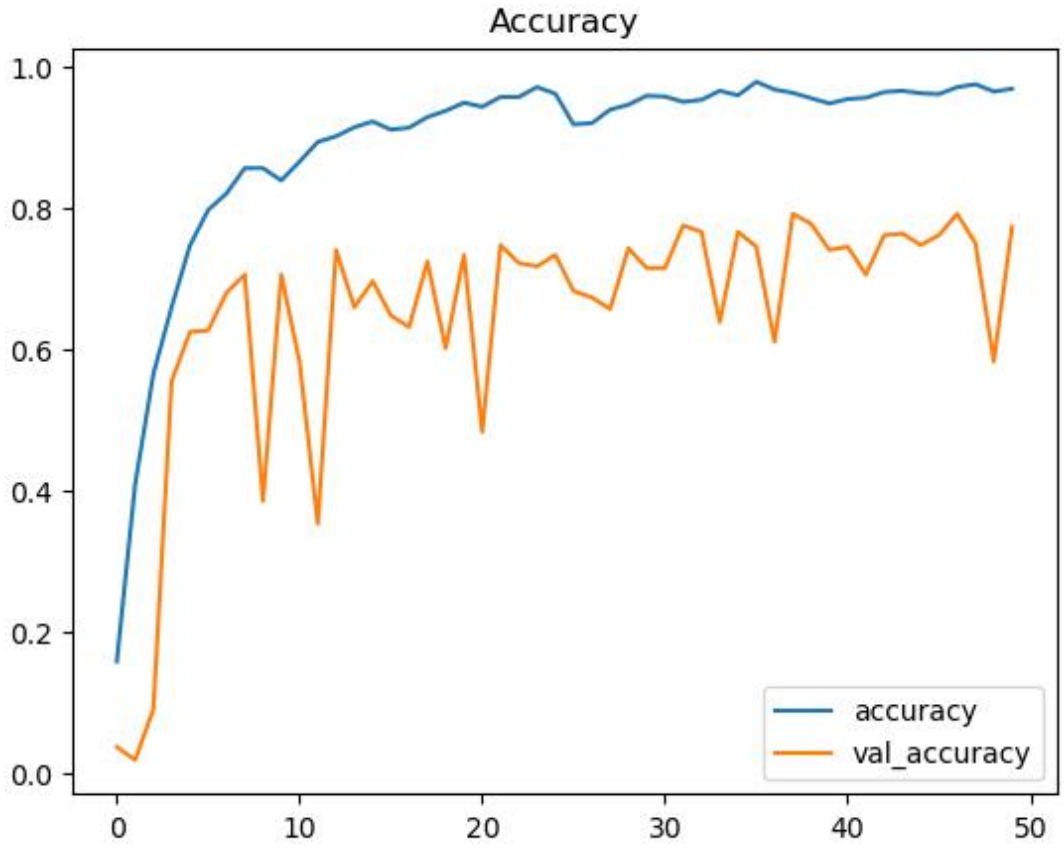
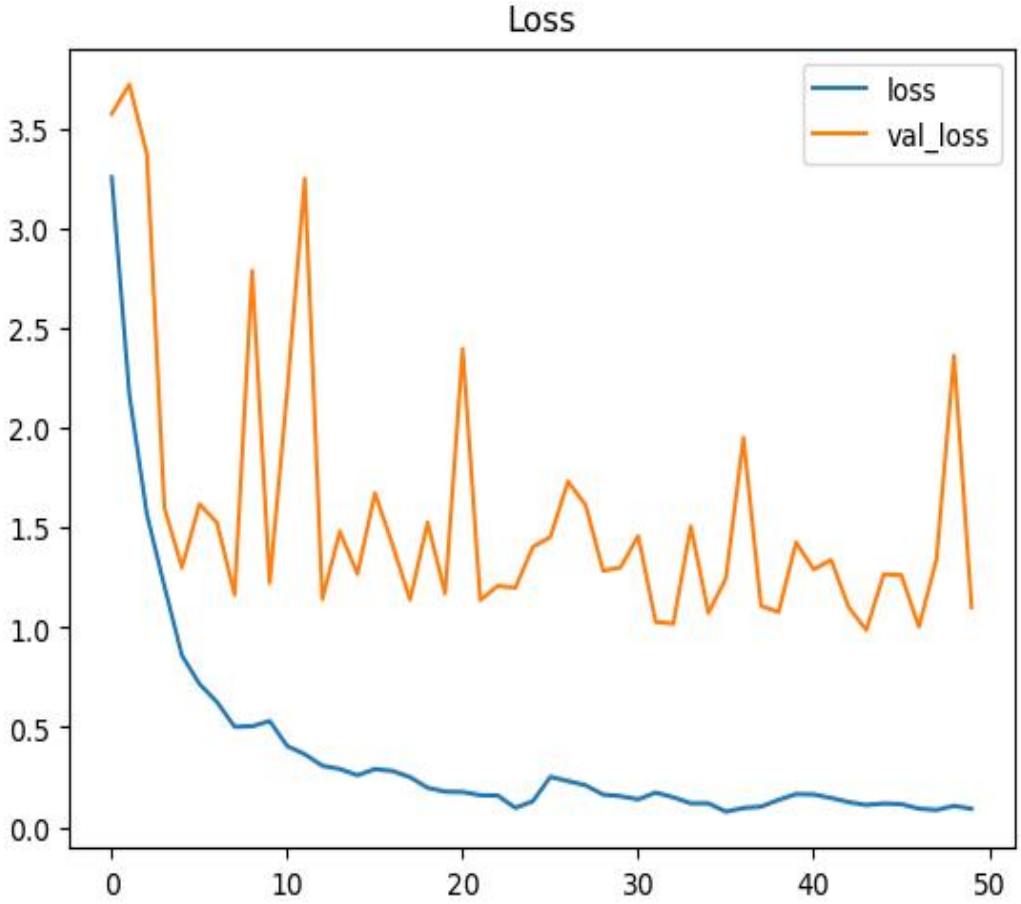


Fig 6.1 Model Accuracy

Snapshot 2: Model Loss

Model loss-Loss Fig 6.2 is the penalty for a bad prediction. That is, loss is a number indicating how bad the model's prediction was on a single example. If the model's prediction is perfect, the loss is zero; otherwise, the loss is greater. The goal of training a model is to find a set of weights and biases that have low loss, on average, across all examples.



Accuracy: 0.75918
Precision: 0.78247
Recall: 0.75918
F1 Score: 0.75419
Cohen Kappa Score: 0.7512

Fig 6.2 Model loss

Snapshot 3: Precision, recall, f1-score and support

From the above we can see the model accuracy & model loss of an individual subtype of cancer in a graph format. The snapshot 3 of Fig 6.3 shows us how accurately the model predicts the subtype. It is represented by a blue line. The orange line represents Val accuracy i.e. nothing but the accuracy of the predictions of a randomly separated model

predicts the subtype. It is represented by a blue line. The orange line represents Val accuracy i.e. nothing but the accuracy of the predictions of a randomly separated validation set after each training period. Our model shows an accuracy of 75.9% for prediction of an individual subtype of cancer.

Snapshot 3 shows us Precision, Recall and F1-Score for each of the 33 Cancer classes. Here,Precision is the ability of a classification model to return only the data points in a class. Recall is the ability of a classification model to identify all data points in a relevant class. F1 score is a single metric that combines recall and precision using the harmonic mean.
The overall accuracy of our model is 76%.

	precision	recall	f1-score	support
ACC	0.67	0.40	0.50	20
BLCA	0.88	0.83	0.86	36
BRCA	0.96	0.92	0.94	25
CESC	0.94	0.57	0.71	30
CHOL	0.67	0.89	0.76	9
COAD	0.79	0.79	0.79	33
DLBC	0.57	0.31	0.40	13
ESCA	0.83	0.34	0.49	29
GBM	0.93	0.47	0.62	30
HNSC	0.90	0.73	0.81	26
KICH	0.85	0.96	0.90	23
KIRC	0.90	0.84	0.87	32
KIRP	0.84	0.82	0.83	33
LAML	0.74	0.71	0.73	28
LGG	0.83	0.80	0.81	30
LIHC	0.73	0.47	0.57	34
LUAD	0.93	0.69	0.79	36
LUSC	0.79	0.76	0.78	25
Meso	0.61	0.61	0.61	23
OV	0.57	0.83	0.68	24
PAAD	0.60	0.84	0.70	32
PCPG	0.59	0.74	0.66	27
PRAD	0.86	0.83	0.85	30
READ	0.91	0.81	0.86	37
SARC	0.69	0.97	0.80	34
SKCM	0.62	0.94	0.75	32
STAD	0.71	0.74	0.73	27
TGCT	1.00	0.93	0.96	29
THCA	0.77	0.74	0.75	27
THYM	0.60	0.97	0.74	32
UCEC	0.86	0.88	0.87	41
UCS	0.56	0.64	0.60	14
UVM	0.63	0.88	0.73	25

accuracy			0.76	926
macro avg	0.77	0.75	0.74	926
weighted avg	0.78	0.76	0.75	926

Fig 6.3 precision, recall ,f1-score and support

Snapshot 4: Confusion Matrix

A confusion matrix Fig 6.4 is a technique for summarizing the performance of a classification algorithm. Classification accuracy alone can be misleading if you have an unequal number of observations in each class or if you have more than two classes in your dataset.Calculating a confusion matrix can give you a better idea of what your classification model is getting right and what types of errors it is making. Through the confusion matrix we can understand that our model makes prediction of subtypes of cancer with good accuracy and minimal amount of errors.

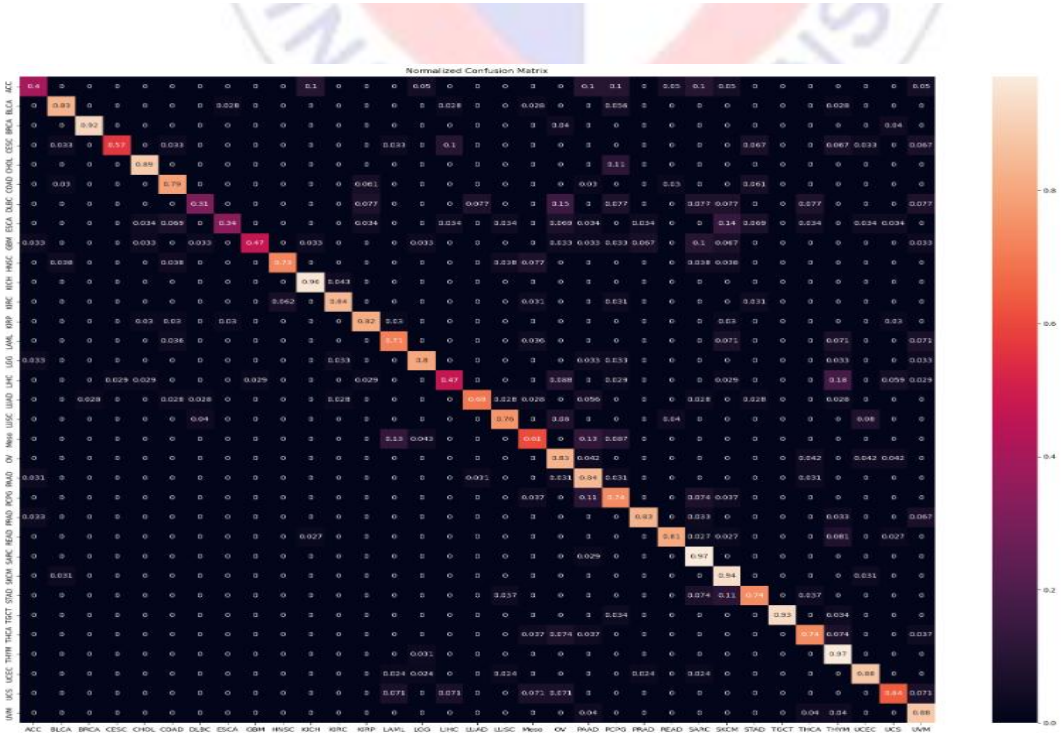


Fig 6.4 Confusion matrix

Chapter 7

Conclusion and future scope

Identification of cancer subtypes is of great importance to cancer diagnosis and therapy. Many approaches were proposed to integrate multi-sources data to identify cancer subtypes in recent years, such as iCluster [11], SNF [8], WSNF [5] and ANF [16], etc. SNF, on behalf of the recent integration methods, is an effective and efficient method that fuses multi-sources data according to similarity networks between samples. It can discover cancer subtypes with different survival patterns by integrating multi-sources data. In view of SNF does not consider the importance of features in similarity fusion, WSNF predicts the importance of features by incorporating mRNA-TF-miRNA regulatory network, and then predict the subtypes based on the integrative similarity information between samples by using SNF framework. It is proved that the heterogeneous regulatory network information is useful to assist similarity estimation between samples. In this study, we consider not only the regulatory associations between features in mRNA-TF-miRNA regulatory network, but also the weights of different data-views. Comparing with WSNF predicts the feature weights directly, CNN extracts multiple expression features for each genomic feature based on the heterogeneous network and uses RV2 matrix correlation method to predict the similarity information between samples. In fact, the extracted features also include the feature importance information in high-dimensionality. In addition, CNN also considers the weights of different data-views in data integration according to manually defining the weight parameters in algorithm. This strategy is more robust to make the method to work on different cancer data.

Although the proposed method detected more significant subtypes in survival estimation on the two cancer datasets, the robustness of the method still need to be further improved in future. Since we predict the similarity information based on the correlation version distance, good data features and data noise processing are important to similarity prediction. The performance of the proposed method hopes to be improved by considering more data information. On the one hand, in this study, we currently based on the datasets in [5], and the network only includes mRNA, TF and miRNA features. More accurate and complete heterogeneous networks hope to provide more comprehensive transcriptome expression.

perform dimension reduction in feature integration at present, which is a linear model in feature embedding. More robust and complex feature-embedding methods may hope to be used to improve the quality of the learned features. We plan to extend and improve the method in future.

In conclusion, we proposed a new model, CNN, to integrate multiple types of transcriptome expression data and heterogeneous biological network to identify cancer subtypes. Tests on TCGA BRCA and GBM datasets demonstrated that the proposed method obtained more significant subtypes, which shown different survival patterns. We hope it could be a useful approach to facilitate the cancer disease analyses in future.

7.1 Major contributions

Subtypes of tumors differ in their imaging characteristics and patterns of metastatic spread. Awareness of these differences will help radiologists to predict tumor subtype at the time of diagnosis and recommend appropriate imaging modalities for additional imaging workup. Likewise, knowledge about patterns of spread will enable radiologists to create additional specific search patterns to identify early and typical metastases. This added skillset will enhance the ability of radiologists to participate in comprehensive cancer care.

Personalized Treatment: Cancer subtyping allows for the identification of distinct molecular and genetic features within a specific cancer type. This knowledge enables personalized treatment strategies tailored to the individual patient. For example, certain subtypes of breast cancer may respond better to specific targeted therapies, and accurately classifying the subtype can guide treatment decisions, leading to improved patient outcomes.

Prognostic Indicators: Cancer subtyping provides valuable prognostic information, allowing clinicians to predict the likely course of the disease and determine the optimal treatment approach. Subtype-specific prognostic indicators, such as gene expression patterns or mutation profiles, help identify patients who may require more aggressive treatment or those who have a better prognosis and may require less intensive therapy.

Precision Medicine: Cancer subtype classification has paved the way for precision medicine, where treatments are tailored to the specific genetic alterations or molecular characteristics of an individual's cancer. By identifying the subtype, clinicians can match

patients with targeted therapies that have shown efficacy against that particular subtype, resulting in more effective treatments and minimizing unnecessary side effects.

Biomarker Discovery: Subtyping cancer has led to the discovery of novel biomarkers that can aid in early detection, diagnosis, and monitoring of the disease. By analyzing the molecular signatures associated with different subtypes, researchers have identified specific biomarkers that can serve as diagnostic tools or indicators of treatment response. These biomarkers have the potential to revolutionize cancer screening and monitoring, leading to earlier detection and intervention.

Research Insights: Cancer subtype classification has provided invaluable insights into the underlying biology of various cancer types. By dissecting the molecular and genetic heterogeneity within a cancer type, researchers can uncover distinct pathways, mechanisms, and driver alterations associated with each subtype. This knowledge drives further research and facilitates the development of new therapeutic targets and innovative treatment strategies.

Clinical Trial Design: Cancer subtyping plays a crucial role in clinical trial design and patient selection. By considering the molecular characteristics of a cancer subtype, researchers can design clinical trials to test targeted therapies specifically for that subtype. This approach increases the likelihood of identifying effective treatments and expedites the development of new therapies tailored to specific subtypes.

7.2 Future Scope

Accuracy: CNNs have already demonstrated impressive accuracy in cancer subtype prediction. However, there is still room for improvement, especially in dealing with challenging cases where inter-class variations are subtle. Future research can focus on refining CNN architectures, exploring novel network designs, and leveraging advanced regularization techniques to enhance accuracy.

Multi-Modal Integration: Cancer diagnosis often involves multiple imaging modalities, such as histopathology, radiology, and molecular imaging. Integrating information from diverse

modalities can provide a more comprehensive understanding of the disease. Future work can investigate CNN-based frameworks that can effectively fuse and learn from multi-modal data, enabling more accurate and robust cancer subtype prediction.

Transfer Learning and Pre-training: CNNs require large labeled datasets for training, which may not always be available, especially for rare cancer subtypes. Transfer learning techniques, where pre-trained models are fine-tuned on a smaller dataset, can alleviate the data scarcity problem. Future research can explore the application of transfer learning and pre-training strategies specifically tailored for cancer subtype prediction, enabling better performance with limited labeled data.

Interpretability and Explainability: CNNs are often referred to as "black box" models due to their complex internal workings, making it challenging to interpret their predictions. As cancer subtype prediction has significant clinical implications, there is a growing need for interpretable and explainable models. Future efforts can focus on developing CNN architectures that can provide insights into the decision-making process, enabling clinicians to understand and trust the model's predictions.

Integration with Genomic and Molecular Data: Genomic and molecular profiling has revolutionized cancer research and treatment. Integrating CNN-based subtype prediction models with genomic and molecular data can provide a more comprehensive understanding of cancer biology. Future research can explore ways to incorporate genomic and molecular features into CNN architectures, allowing for more accurate and personalized cancer subtype prediction.

References

[1] Li, Y., Kang, K., Krahn, J. M., Croutwater, N., Lee, K., Umbach, D. M., & Li, L.“ A comprehensive genomic pan-cancer classification using The Cancer Genome Atlas gene expression data”. BMC genomics, 18(1), 1-13 ,2020.

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

[3] Rocha, D., García, I. A., González Montoro, A., Llera, A., Prato, L., Girotti, M. R., & Fernández, E. A. “Pan-Cancer Molecular Patterns and Biological Implications Associated with a Tumor-Specific Molecular Signature”. Cells, 10(1), 45,2020.

[4] López-García, G., Jerez, J. M., Franco, L., & Veredas, F. J. “Transfer learning with convolutional neural networks for cancer survival prediction using gene-expression data.” PloS one, 15(3), e0230536,2019.

[5] Chaudhari, P., Agrawal, H., & Kotecha, K. “Data augmentation using MG-GAN for improved cancer classification on gene expression data.” Soft Computing, 24(15), 11381-11391,2020.

[6] Kim, B. H., Yu, K., & Lee, P. C. “Cancer classification of single-cell gene expression data by neural network. Bioinformatics”, 36(5), 1360-1366,2018.

[7] Rao, S. Mitos-rcnn: “A novel approach to mitotic Fig detection in breast cancer histopathology images using region based convolutional neural networks.” arXiv preprint arXiv:1807.01788 ,2018.

[8] Bejnordi, B. E. et al. “Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer”. Jama 318, 2199–2210 ,2019.

[9] Bándi, Petal. “From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge”. IEEE Transactions on Med. Imaging Cancer treatment recommendation system

[10] Litjens, G. et al. 1399 “h&e-stained sentinel lymph node sections of breast cancer patients: the camelyon dataset.” GigaScience 7, giy065 2018.

[11] Aresta, G. et al. “Bach: Grand challenge on breast cancer histology images.” arXivpreprintarXiv:1808.04277 2018.

[12] Gouda I Salama, M Abdelhalim, and MagdyAbdelghanyZeid. “Breast cancer diagnosis on three different datasets using multiclassifiers.” Breast Cancer (WDBC) 32, 569 ,2020, 2.

[13] William H Wolberg, W Nick Street, and Olvi L Mangasarian. 1992. Breast cancer Wisconsin (diagnostic) data set. “UCI Machine Learning Repository” [http://archive. ics. uci. edu/ml/], 2019.

[14] Abien Fred Agarap. 2017. “A Neural Network Architecture Combining Gated Recurrent Unit(GRU) and Support Vector Machine (SVM) for Intrusion Detection in Network Traffic Data”. arXiv preprint arXiv:1709.03082 (2020).

[15] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, HolgerSchwenk, and Yoshua Bengio, (2014), “Learning phrase representations using RNN encoder-decoder for statistical machine translation”, arXiv preprint arXiv:1406.1078 (2021).

[16] Gönen, M.; Alpaydın, E. Multiple kernel learning algorithms. J. Mach. Learn. Res.2018, 12, 2211–2268.

[17] Chang Ming, Valeria Viassolo, Nicole Probst-Hensch, Pierre O Chappuis, Ivo D. Dinov & Maria C. Katapodi, Machine learning techniques for personalized breast cancer risk prediction: comparison with the BCRAT and BOADICEA models, Breast Cancer Research volume 21, Article number: 75

[18] Ch. Shravya, K. Pravalika, Shaik Subhani, “Prediction of Breast Cancer Using Supervised Machine Learning Techniques,” International Journal of Innovative Technology and Exploring Engineering , Volume-8 Issue-6, April 2019.

[19] MandeepRana, PoojaChandorkar, AlishibaDsouza, "Breast cancer diagnosis and recurrence prediction using machine learning techniques", International Journal of Research in Engineering and Technology Volume 04, Issue 04, April 2019.

[20] Hiba Asri,Hajar Mousannif,Hassan Al "Moatassime,Thomas Noel, Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis," The 6th International Symposium on Frontiers in Ambient and Mobile Systems (FAMS 2016), Procedia Computer Science 83 ,2020, 1064 – 1069.

[21] Habib Dhahri,Eslam Al Maghayreh, Awais Mahmood,Wail Elkilani,and Mohammed Faisal Nagi, "Automated Breast Cancer Diagnosis Based on Machine Learning Algorithms, Journal of HealthCare Engineering" , Volume 2019 .

[22] Hiba Asri,Hajar Mousannif,Hassan Al Moatassime,Thomas Noel, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis," The 6th International Symposium on Frontiers in Ambient and Mobile Systems (FAMS 2016), Procedia Computer Science 83 (2016) 1064 – 1069.

Appendix

RV Institute of Technology & Management

Chaithanya Layout, 8th Phase, JP Nagar, Bangalore, Karnataka – 560076 Recognised by AICTE, New Delhi | Affiliated to VTU, Belagavi

CERTIFICATE

OF PARTICIPATION

This is to certify that

Naresh Santosh Shet

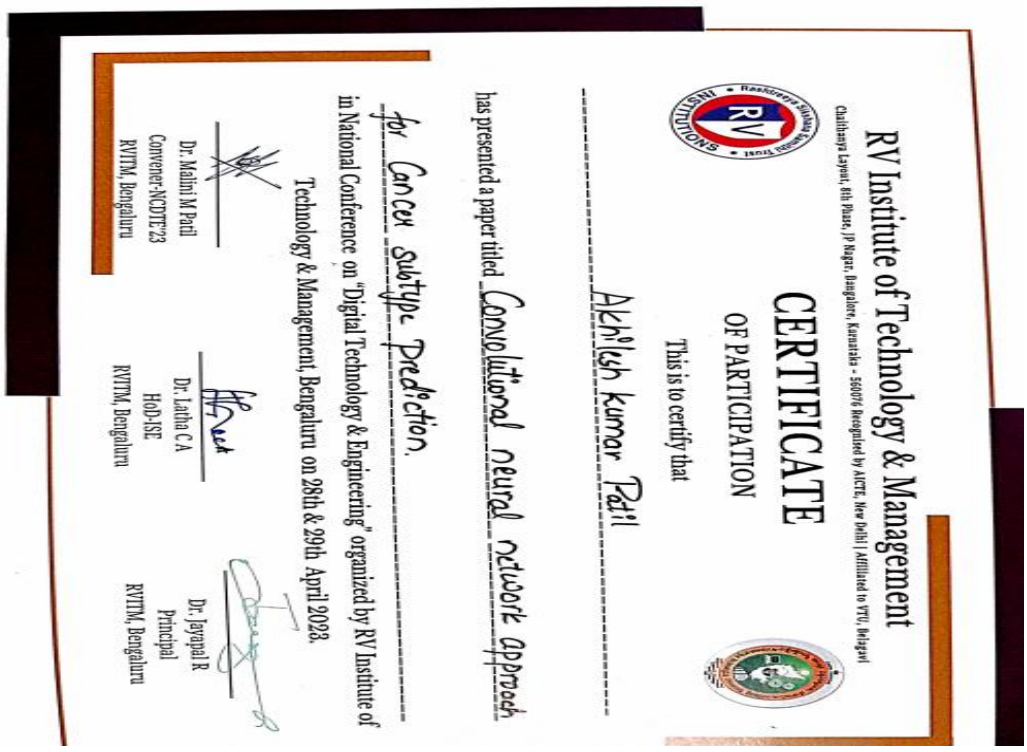
has presented a paper titled Convolutional Neural Network Approach for Cancer subtype Prediction

in National Conference on "Digital Technology & Engineering" organized by RV Institute of Technology & Management, Bengaluru on 28th & 29th April 2023.

Dr. Malini M Patil
Convener-NCDTE'23
RVITM, Bengaluru

Dr. Latha C A
HoD-ISE
RVITM, Bengaluru

Dr. Jayapal R
Principal
RVITM, Bengaluru



RV Institute of Technology & Management

Chaithanya Layout, 8th Phase, JP Nagar, Bangalore, Karnataka – 560076 Recognised by AICTE, New Delhi | Affiliated to VTU, Belagavi

CERTIFICATE

OF PARTICIPATION

This is to certify that

Akhilesh kumar Patil

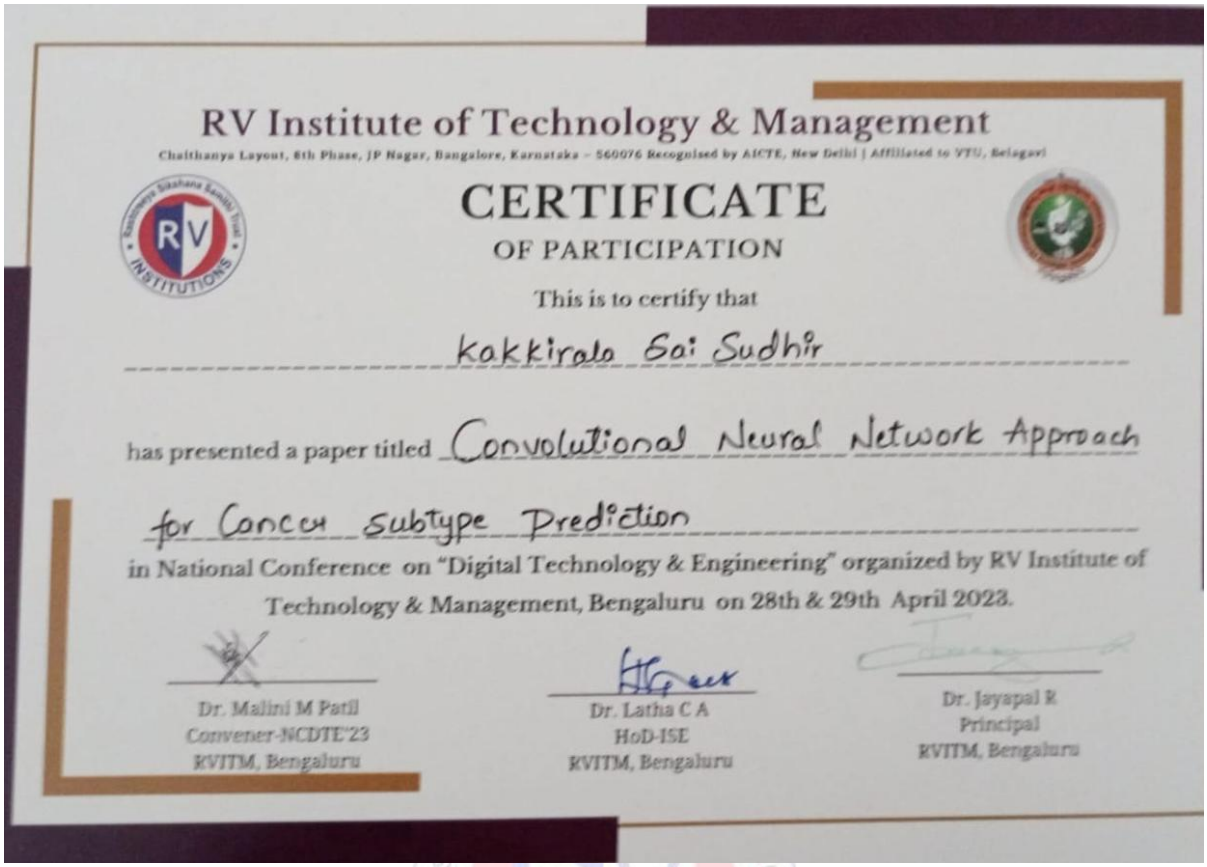
has presented a paper titled Convolutional neural network approach for Cancer subtype prediction.

in National Conference on "Digital Technology & Engineering" organized by RV Institute of Technology & Management, Bengaluru on 28th & 29th April 2023.

Dr. Malini M Patil
Convener-NCDTE'23
RVITM, Bengaluru

Dr. Latha C A
HoD-ISE
RVITM, Bengaluru

Dr. Jayapal R
Principal
RVITM, Bengaluru



RV Institute of Technology & Management

Chaithanya Layout, 8th Phase, JP Nagar, Bangalore, Karnataka – 560076 Recognised by AICTE, New Delhi | Affiliated to VTU, Belagavi

CERTIFICATE

OF PARTICIPATION

This is to certify that

Kakkirala Sai Sudhir

has presented a paper titled Convolutional Neural Network Approach for Cancer Subtype Prediction

in National Conference on "Digital Technology & Engineering" organized by RV Institute of Technology & Management, Bengaluru on 28th & 29th April 2023.

Dr. Malini M Patil
Convener-NCDTE'23
RVITM, Bengaluru

Dr. Latha C A
HoD-ISE
RVITM, Bengaluru

Dr. Jayapal R
Principal
RVITM, Bengaluru