IMPLEMENTATION OF BERT BASED MACHINE LEARNING MODEL TO EXTRACT CANCER –MIRNA RELATIONSHIP FROM RESEARCH LITERATURE

by

Arunprasad Sundharam

December, 2021

Director of Thesis: Qin Ding, PhD

Major Department: Computer Science

I. Abstract:

In the world today, text mining is a widely popular and growing branch of Information technology,in which we extract useful information out of the given pile of text data. There are thousands of research papers in medical science pertaining to the study of how microRNAs (miRNAs) can assist or impede the development of varioustypes of cancers. mirCancer is a repository which provides the details of this cancer-miRNA association by analyzing 6500+ research papers using text mining techniques. It would be helpful to create a machine learning model which can analyze the title and abstract content of the research papers and extract the cancer-miRNA association details if it is available in the given text. In this thesis work, we are proposing a solution for creating a machine learning model using the open source NLP framework – **BERT**, provided by Googlewhich can identify the cancer-miRNA relationship in the given abstract text content. Bert is a deep learning model which is pretrained on Wikipedia text corpse and has built-in knowledge on the usage of english language. As part of this work, we have designed and implemented a machine learning model using Bert framework along with preparation of the dataset required to train the model in the task of identifying cancer-miRNA relationship from the given text. The machine learning model developed in this thesis work performed with an overall accuracy of 90.3% in retrieving the required information from the research papers of the test dataset and hence it can be leveraged to review the results of the existing mircancer text mining implementation.

TITLE

IMPLEMENTATION OF BERT BASED MACHINE LEARNING MODEL TO EXTRACT CANCER –MIRNA RELATIONSHIP FROM RESEARCH LITERATURE

A Thesis

Presented to The Faculty of the Department of Computer Science

East Carolina University

In Partial Fulfillment of the Requirements for the Degree

Master of Science in Computer Science

by

Arunprasad Sundharam

December, 2021

IMPLEMENTATION OF BERT BASED MACHINE LEARNING MODEL TO EXTRACT CANCER –MIRNA RELATIONSHIP FROM RESEARCH LITERATURE

By

Arunprasad Sundharam

APPROVED BY:

DIRECTOR OF THESIS: _____

Qin Ding, PhD

COMMITTEE MEMBER: _____

Rui Wu, PhD

COMMITTEE MEMBER: _____

Venkat N. Gudivada, PhD

CHAIR OF THE DEPARTMENT
OF COMPUTER SCIENCE: _____

Venkat N. Gudivada, PhD

DEAN OF THE
GRADUATE SCHOOL: _____

Paul Gemperline, PhD

DEDICATION

I would like to dedicate the successful completion of this thesis to my beloved wife who was able to understand my career ambitions, and has been a great moral support in my journey to pursue this exciting learning and research opportunity as part of ongoing my masters degree, which I believe will help me to succeed in my professional career further.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1 - Role of microRNA in the development of cancer

1.1 Introduction to microRNA

MicroRNAs (miRNAs) are a class of non-coding RNAswith an average of 22 nucleotides in length and play important roles in regulating gene expression. The majority of miRNAs are transcribed from DNA sequences into primary miRNAs and processed into precursor miRNAs, and finally mature miRNAs. In most cases, miRNAs interact with the 3′ untranslated region (3′ UTR) of target mRNAs to induce mRNA degradation and translational repression. However, interaction of miRNAs with other regions, including the 5′ UTR, coding sequence, and gene promoters, have also been reported. Under certain conditions, miRNAs can also activate translation or regulate transcription. The interaction of miRNAs with their target genes is dynamic and dependent on many factors, such as subcellular location of miRNAs, the abundancy of miRNAs and target mRNAs, and the affinity of miRNA-mRNA interactions. miRNAs can be secreted into extracellular fluids and transported to target cells via vesicles, such as exosomes, or by binding to proteins, including Argonautes. Extracellular miRNAs function as chemical messengers to mediate cell-cell communication. MicroRNAs (miRNAs) are involved in the regulation of a variety of biological and pathological processes, including the formation and development of cancer.

1.2 Background of microRNA

The discovery of the first microRNA (miRNA), lin-4, in 1993 by the Ambros and Ruvkun groups in Caenorhabditis elegans has revolutionized the field of molecular biology [9]. Years before, lin-4 was characterized by the Horvitz's lab as one of the genes that regulate temporal development of C. elegans larvae. Later in 1987, the same group found that a mutation in lin-4 had an opposite phenotype to a mutation in another gene, lin-14, yet a lin-14 suppressor mutation in a null-lin-4 line was wildtype. Both Ambros and Ruvkun continued to study lin-4 and lin-14 after leaving the Horvitz's lab, only to discover later that lin-4 was not a protein-coding RNA but indeed a small non-coding RNA. They also found that lin-14 was post-transcriptionally downregulated through its 3′ untranslated region (UTR) and that lin-4 had a complementary sequence to that of the 3′ UTR of lin-14 . Therefore, they proposed that lin-4 regulates lin-14 at the post-transcriptional level [9]. Since then, miRNAs have been detected in all

animal model systems and some were shown to be highly conserved across species. New miRNAs are still being discovered and their roles in gene regulation are well recognized.

Most miRNAs are transcribed from DNA sequences into primary miRNAs (pri-miRNAs) and processed into precursor miRNAs (pre-miRNAs) and mature miRNAs. In most cases, miRNAs interact with the 3′ UTR of target mRNAs to suppress expression. However, interaction of miRNAs with other regions, including the 5′ UTR, coding sequence, and gene promoters, have also been reported. Furthermore, miRNAs have been shown to activate gene expression under certain conditions. Recent studies have suggested that miRNAs are shuttled between different subcellular compartments to control the rate of translation, and even transcription.

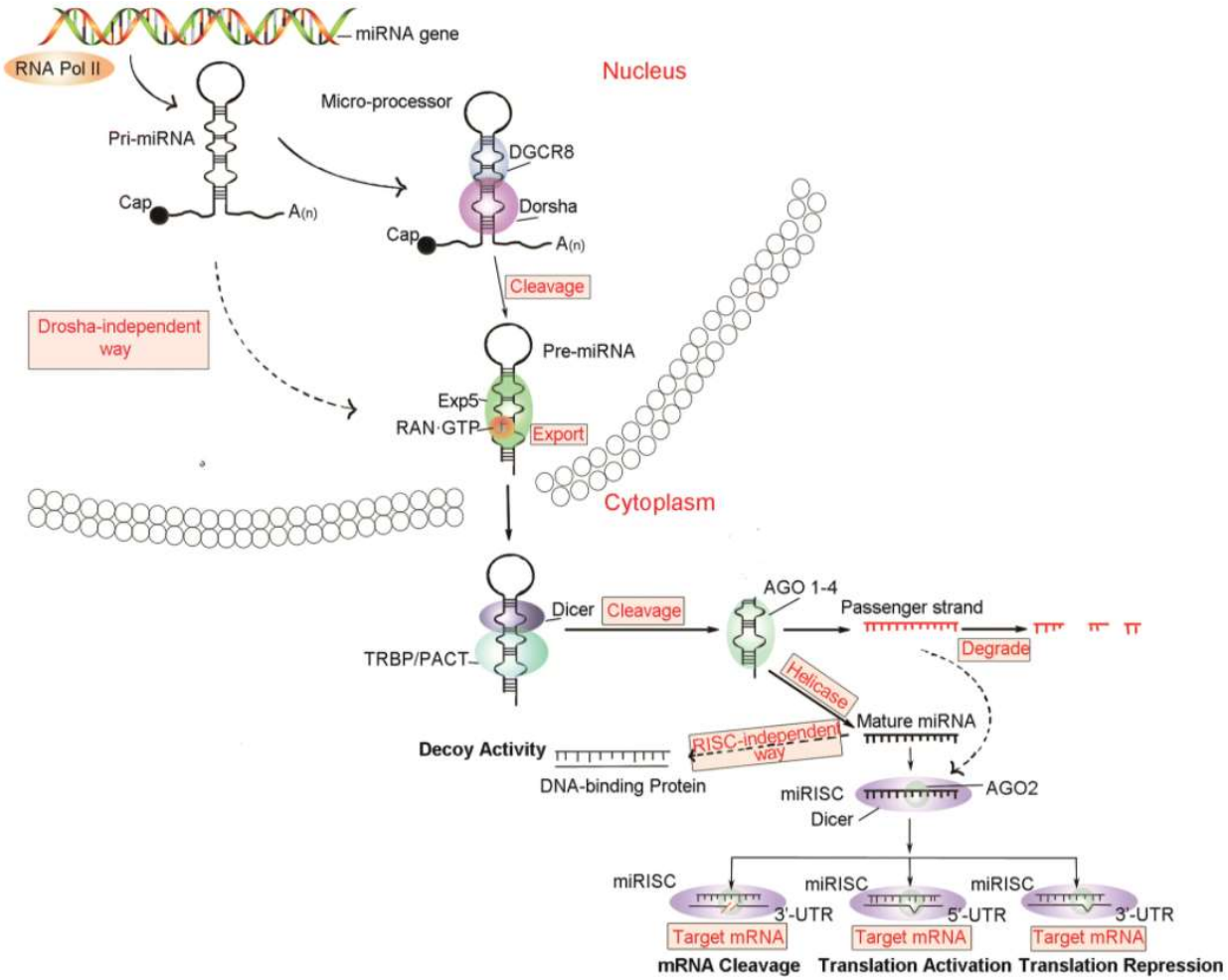
1.3 Biogenesis of microRNAs

The biogenesis of miRNAs in animal cells and the mechanisms of regulation of their target gene expression are shown in Figure No 1- Biogenesis of miRNAs. In simple terms, this process can be divided into the following steps

1. The miRNA gene is transcribed into primary miRNA (pri-miRNA) by RNA polymeraseII(RNA polII) in the nucleus.
2. pri-miRNA is processed by the Drosha/DGCR8 complex to release the intermediate precursor miRNA (pre-miRNA), which is approximately 70 nt with a stem loop structure and a 2 nt overhang at the 3′-end.
3. pre-miRNA binds to the Exp5/Ran-GTP complex, which allows for its transport into the cytoplasm.
4. The pre-miRNA is then processed into double-stranded RNA by the Dicer/TRBP/PACT complex in the cytoplasm.
5. The miRNA-duplex is unwound into single strands by the action of helicase. Under normal circumstances, the RNA strand with lower stability at the 5′-end will be integrated into the RNA-induced silencing complex (RISC) and become a mature miRNA, and the strand with higher stability at the 5′-end will be degraded.

6. miRNA-induced silencing complex (miRISC) will bind to the 3′-untranslated regions (UTR) of the target mRNA, thus inhibiting its translation.

Figure No 1- Biogenesis of miRNAs



The mechanisms of microRNA biogenesis and its regulation of gene expression. The solid arrows represents the classical pathway, the dotted arrows represents the non-classical pathway

(Picture added from courtesy of external references on microRNA. Please refer item no-6 in the references section)

In plant cells, miRISC will degrade its target mRNA, and the biogenesis of miRNAs is slightly different from that in animal cells.Existing research shows that this classical processing and functioning pathway has some exceptions. For example, in step (2) of its biosynthesis process, the pri-miRNA can also be processed into pre-miRNA in a Drosha-independent way. In step (5), the two strands may be

randomly integrated into RISC, or they could bind to mRNA in RISC-independent manner. In step (6), some miRISC can bind to the 5′-UTR of mRNA, upregulating its translation.

Through the above approach, miRNAs regulate about 30% of human genes. Half of these genes are tumor-associated. The deregulation of miRNAs in tumor cells suggests that they have modulatory effects on tumor development. In fact, some miRNAs may act as tumor genes, and others can act as tumor suppressor genes. Interestingly, some can act as both, depending on the tissue where they are being expressed. In recent years, a large number of studies have indicated that many of the genes regulated by miRNAs are related to the response of tumor cells to chemotherapeutic agents.

1.4 Association of MIRNA on Cancer

MiRNAs are involved in the regulation of a variety of biological processes, such as cell cycle, differentiation, proliferation, apoptosis, stress tolerance, energy metabolism, and immune response,and are critical for normal animal development. In addition, miRNAs are secreted into extracellular fluids. Extracellular miRNAs have been widely reported as potential biomarkers for a variety of diseases and they also serve as signaling molecules to mediate cell-cell communications. Aberrant expression of miRNAs is associated with many human diseases including cancer.Cancer is a leading cause of death worldwide. It is estimated that cancer caused 7.6 million deaths in 2008, about 13% of all deaths globally. By 2010, cancer surpassed heart disease as the top killer in USA for the first time. There are many possible carcinogens including tobacco, radiation, chemicals, environmental toxins, viruses and genetic problems. However, the causes of many cancers remain unknown. Cancer involves unregulated cell growth, thereby invading nearby parts of body during development. Early diagnosis before proliferation usually makes a difference in treatment and survival rate. The fact that miRNA expression levels vary significantly between normal cells and cancer cells suggests that miRNA might be associated with cancer development and potentially could be used for cancer diagnosis or even treatment. Even though it is uncertain whether cancer is a cause or consequence of deviant miRNA expression, miRNA fingerprints are found in all types of analysed cancers, such as lung cancer, breast cancer, cervical cancer and lymphoblastic leukemia.

Chapter 2 - Text Mining in Biological domain

2.1 Research Interest on cancer-miRNA association

There are thousands of research papers published about the association of miRNAs in various types of cancer.As the research interests on this topic keeps growing, there is also need to consolidate the findings of these research papers for future works. Numerous databases have been created to document miRNA functionalities either from computational predictions or from experimental results. Although computational target prediction methods are fast, experimental validation of miRNA functionalities is also needed. The significant increase in validation experiments raises the need for having a database to store these results in some uniform way.

However, comparing to databases providing computationally predicted miRNA functions, databases storing experimental miRNA targets are rare. There are databases which provide the details of the miRNA association on various diseases. Rapid increase in the number of miRNA-related publications makes the manual collection more and more difficult.

2.2 Text Mining of Biological Domain text

Text mining is the process in which useful information is extracted from text using computational approaches or tools. Unlike its application in other fields, accurate biomedical text mining remains an open problem as a result of specialized and complex vocabularies. There are three commonly used text-mining approaches in biomedical realm:

(i) co-occurrence- based approach, normally easier to build while the other two provide better accuracy

(ii) rule-based approach, which keeps a set of rules that usually take significant amount of time to develop

(iii) machine learning, where the required training data are usually expensive

2.3 miRCancerRepository

miRCancer is a respository (hosted by East Carolina University) where medical researchers can quickly access the microRNA–cancer associations details determined from the experimental results which are published from 6400+ research papers [10]. This repository was developed using rule-based textmining approaches to extract miRNA and cancer association and store them in a database [1]. All the discovered associations have been manually confirmed after automatic extraction. miRCancer initially in 2013 documented 878 relationships between 236 microRNAs and 479 human cancers through the processing of 426000 published articles from PubMed. Over the years, the repository has grown significantly and currently documents 9080 relationships between 1037 microRNAs and 131 human cancers. In this Thesis work, we wanted to pursue an alternate machine learning based approach to extract the cancer-miRNA relationship from the given abstract text using the open source Bert framework.

2.4 Bert Model Introduction

Bert is the acronym for **Bi-Directional Encode-Decoder Representation from Transformers.** Bert is a deep learning model which is pretrained on Wikipedia text corpse and has built-in knowledge on the usage of English language.It can be used for downstream NLP tasks like question answering, sentence classification. BERT is the first unsupervised, deeply bidirectional system for pre-training Natural Language Processing (NLP).Unsupervised means that BERT was trained using only a plain text corpus, which is important because an enormous amount of plain text data is publicly available on the web in many languages.

2.5 Different Flavors of Bert Model

BERT is a multi-layer bidirectional Transformer encoder. Bert model architecture consists of

(I) 12-layer to 24-layer Transformers

(II) 12 to 16 attention heads

(III) 110 to 340 million parameters.

Various flavors of Bert are available depending on the number of Transformers, attention heads and parameters as listed below:

- BERT-Large, Uncased (Whole Word Masking): 24-layer, 1024-hidden, 16-heads, 340M parameters
- BERT-Large, Cased (Whole Word Masking): 24-layer, 1024-hidden, 16-heads, 340M parameters
- BERT-Base, Uncased: 12-layer, 768-hidden, 12-heads, 110M parameters
- BERT-Large, Uncased: 24-layer, 1024-hidden, 16-heads, 340M parameters
- BERT-Base, Cased: 12-layer, 768-hidden, 12-heads , 110M parameters
- BERT-Large, Cased: 24-layer, 1024-hidden, 16-heads, 340M parameters
- BERT-Base, Multilingual Cased (New, recommended): 104 languages, 12-layer, 768-hidden, 12-heads, 110M parameters
- BERT-Base, Multilingual Uncased (Orig, not recommended) (Not recommended, use Multilingual Cased instead): 102 languages, 12-layer, 768-hidden, 12-heads, 110M parameters
- BERT-Base, Chinese: Chinese Simplified and Traditional, 12-layer, 768-hidden, 12-heads, 110M parameters

2.6 Using Bert for NLP

Using BERT for NLP consists of two stages, pre-training and fine-tuning.

1. **Pre-training**

Pre-training is fairly expensive (four days on 4 to 16 Cloud TPUs), but is a one-time procedure for each language (current models are English-only, but multilingual models will be released in the near future). Google had released a number of flavors of pre-trained models. Most NLP researchers will never need to pre-train their own model from scratch.

2. Fine-tuning

Fine-tuning is inexpensive which requires mostly 1 hour of training on a single Cloud TPU, or a few hours on a GPU, starting from the exact same pre-trained model. Fine-tuning is required to make BERT adapt to different types of NLP tasks like sentence classification, question answering etc.

Chapter 3 - Proposing a Machine Learning solution

3.1 Problem Analysis

We took inspiration out of the mirCancer repository and started this thesis to extract the cancer-miRNA association details from the abstract text content of the research papers using an alternate machine learning model approach. We determined that the proposed machine learning model should be able to retrieve the below 3 types of information from each abstract text as part of the text extraction process.

1. Type of Cancer
2. Name of the miRNA
3. miRNA Regulation

We analyzed the abstract content of the research papers to understand the following details regarding each Information type

1. How many unique values are possible for the Information Type in the given abstract text?
2. Can the value of the Information Type be retrieved directly from the given abstract text?
3. Can the abstract text contain duplicate values for the Information type?

We prepared the below table with details of our preliminary analysis about each information type. We refined our preliminary analysis further on each information type so that we can design the machine learning appropriately to fetch the required information from the abstract text.

3.1.1 Type of cancer

Eventhough we observed in our analysis that there could be more than one type of cancer mentioned in the abstract text, we determined only 9.32% research papers had more than one cancer type. In this Thesis, we are targeting to retrieve only one cancer type from the given abstract text. If the abstract text contains more than cancer type, the machine learning model which we propose in this thesis work will retrieve only the cancer type with first occurrence in the given abstract text.

As we are capturing only the first occurrence of this information type, we are not also concerned about the other duplicate occurrences of the retrieved value of cancer type.

Table-1: (Analysis of Information Types)

Sno	Information Type	Number of possible target values in the abstract text	Can the target Information be retrieved directly from the abstract text	Comments
1	Type of Cancer	More than one cancer type	Yes	The abstract text can contain information of more than one type of cancer.
2	Name of the miRNA	Multiple distinct miRNAs are possible in the input text content	Yes	The abstract text can contain information of more than one miRNA, associated with the cancer.
3	miRNA Regulation	2 Possible values- 'UP' or 'DOWN'	No	The abstract text contain the relevant sub text which describes how the miRNA regulates the cancer (either UP or DOWN)

3.1.2 Name of the miRNA

We understand there could be multiple distinct miRNA names mentioned in the abstract text. Also the given text can also have duplicate occurences of the same miRNA. We would actually need to capture all the distinct miRNA names mentioned in the input text content.

3.1.3 Cancer-miRNA regulation association

Out of all the distinct miRNAs studied in the given abstract text, there could be one or more miRNA which has been observed to regulate/promote with the specified type of cancer. Also all distinct miRNAs mentioned in the abstract text may not have regulation association with the specified cancer type

in the text.The Cancer-miRNA regulation associationtext retrieved from the abstract content still needs to be infered/ classified as ‘UP’ or ‘DOWN’ regulated.

3.2 Proposed Design of Bert based Machine Learning Model

We are proposing the Machine Learning model, shown in **Figure No 2- Proposed Machine Learning Model-Architecture** as part of this thesis work to retrieve all the 3 required Information types from the abstract text. The given abstract content along with the title of the research, (together we will refer this as **input text content**, hereafter) is provided as input to this Bert Model along with appropriate questions regarding each information type. As mentioned earlier, this machine learning model is designed to extract only mirna association of only one type of cancer text from each given abstract text using this machine learning model. We observed a small percentage (9.32%) of research papers study the mirna association with more than one type of cancer.

The proposed machine learning model consists of 2 separate Bert Models-

1. The First Bert Model is finetuned on Question and Answering task.
2. The Second Bert Model is finetuned on Sentence Classification task.

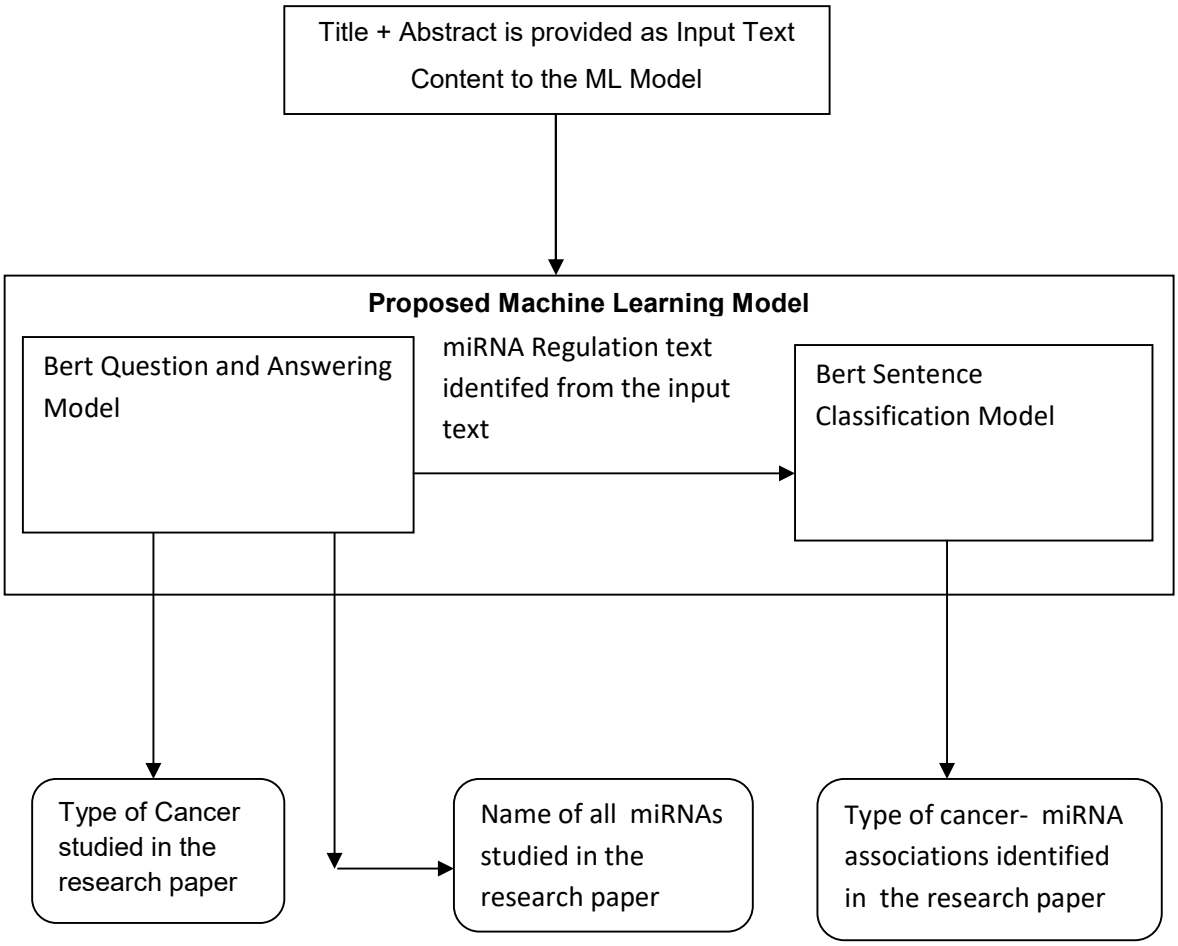
3.2.1 Bert Model finetuned for Question and Answering:

The given input text content is provided as input first into this Bert Model along with appropriate questions regarding each information type. The model will retrieve the appropriate text value for each information type if the model finds suitable answers in the given input text content for the questions related to each information type.

3.2.2 Bert Model finetuned for Sentence Classification:

This bert model is intended to classify the miRNA Regulation text retrieved by the first bert model (finetuned for Question and answering) for the cancer-miRNA regulation association as UP or DOWN regulated.

Figure No 2- Proposed Machine Learning Model-Architecture



Chapter 4 - Setting up the Infrastructure and Corpus Text

4.1 Setting up the Hardware and Software Infrastructure:

Below are the details of the Hardware and Software technology stack used by us for this Thesis work.

4.1.1 Hardware Details:

1. Google CoLab:

Google CoLab subscription was purchased to run the training and validation process on the proposed machine learning model. It costs about 9.99 usd per month for the subscription.

2. Google Cloud Storage console:

We used the google cloud storage console subscription to store all the bertQA and bertSC models that we created as part of the training process. This is a free subscription for 365 days after which we need to upgrade to premium subscription.

3. We used a personal laptop to connect to the above google cloud infrastructure and for carrying out all activities related to this thesis work.

4.1.2 Software Details:

1. Python Programming language –V3.6 with Anaconda.
2. Spyder IDE package for Python.
3. Hugging Face Transformers Python Package-(V3.5.1)
4. Hugging Face Transformers source code from GitHub -(V3.5.1).
5. Oracle Java SE edition(V8)
6. Eclipse photon edition
7. Selenium Java package
8. VBA programming using Microsoft excel
9. Google Chrome browser
10. Adobe pdf acrobat reader
11. Microsoft Excel and Word (V2007)

12. Notepad ++ (V7.6)
13. Online utilities to perform the below functions:
    a. Text to tsv format converter.
    b. Test Regular expression
    c. Format Json
    d. Compare text/json

4.2 Preparing miRNA-Cancer association Corpus Text:

We gathered the details of 6422 research papers related to the study of cancer-miRNA association from the miRCancer repository (mircancer) . We then created an automation script to capture the actual abstract content of all the research papers from the PubMed website. Some 13 research papers did not have abstracts available as these articles were retracted. We split the abstracts of the available 6409 research topics into two set of corpuses.

1. **Training Dataset:**

   One corpus (with 5197 abstracts) was used for creation of training dataset. Out of 5197 abstracts, we observed 188 abstracts had more than 512 tokens (words in Bert vocabulary) limit and hence cannot be further parsed by the Bert model. Also 330 abstracts did not contain the expected cancer-miRNA relationship in the provided text during manual verification. After removing these unparsable abstracts, we finally had a corpus of 4679 abstracts to be used as the training dataset.

2. **Test Dataset:**

   Another corpus (with 1212 abstracts) was used for creation of test dataset. Out of 1212 abstracts, we observed 34 abstracts had more than 512 tokens (words in Bert vocabulary) limit and hence cannot be further parsed by the Bert model. Also 33 abstracts did not contain the expected cancer-miRNA relationship in the provided text during manual verification. After removing these unparsable abstracts, we finally had a corpus of 1145 abstracts to be used as the test dataset.

We created the below 5 csv files to save the details of the abstract content in the corpus text. (listed in APPENDIX A)

1. AbstractDump.csv → This csv file contains details of title of the research paper and input text content of the research paper.
2. cancerDump.csv → This csv file contains details of title of the research paper and type of cancer studied in the research paper
3. miRNADump.csv → This csv file contains details of title of the research paper and all the miRNAs studied in the research paper.
4. miRNARegulationDump.csv → This csv file contains details of title of the research paper and regularion type (UP or DOWN regulated) of each miRNA studied in the research paper.
5. RegulationTextFinal.csv → This csv file contains details of the actual text in the input text content describing if the cancer-miRNA association is UP or DOWN regulated.

As we have 2 set of corpus texts now, we created two versions of the above 5 csv files- one set of csv files used in the creation of the training dataset whereas the other set used for the test dataset.

4.3 Implementing the Proposed Machine Learning Model:

The below activities are carried out in order to implement the proposed machine model

1. Finetune the first bert model for question answering task
    a) Design the training strategy for question answering task to retrieve the correct text for the 3 information types.
    b) Create the training tataset in the SQUAD format expected by bert model for question and answering task
    c) Train the bert model using the created training dataset.
2. Finetune the second bert model for sentence classification task
    a) Create the Training Dataset in the SST-2 format expected by bert model for sentence classification task
    b) Train the bert model using the created training dataset.

3. Develop an automated solution to validate the final actual output values of the machine learning against the expected results.
4. Create the separate set of test dataset to evaluate the machine learning model.
5. Validate the created model against the test dataset using the developed automated validation solution.

Chapter 5 - Setting up and finetuning the Bert- Question Answering Model

5.1 Activities for implementing the Bert-Question Answering Model

Below are the list of tasks performed in order to create the required bert question answering model specified the architecture.

a) Design the training strategy for question answering task to retrieve the correct text for the 3 information types.

b) Create the training dataset in the SQUAD format expected by bert model for question answering task

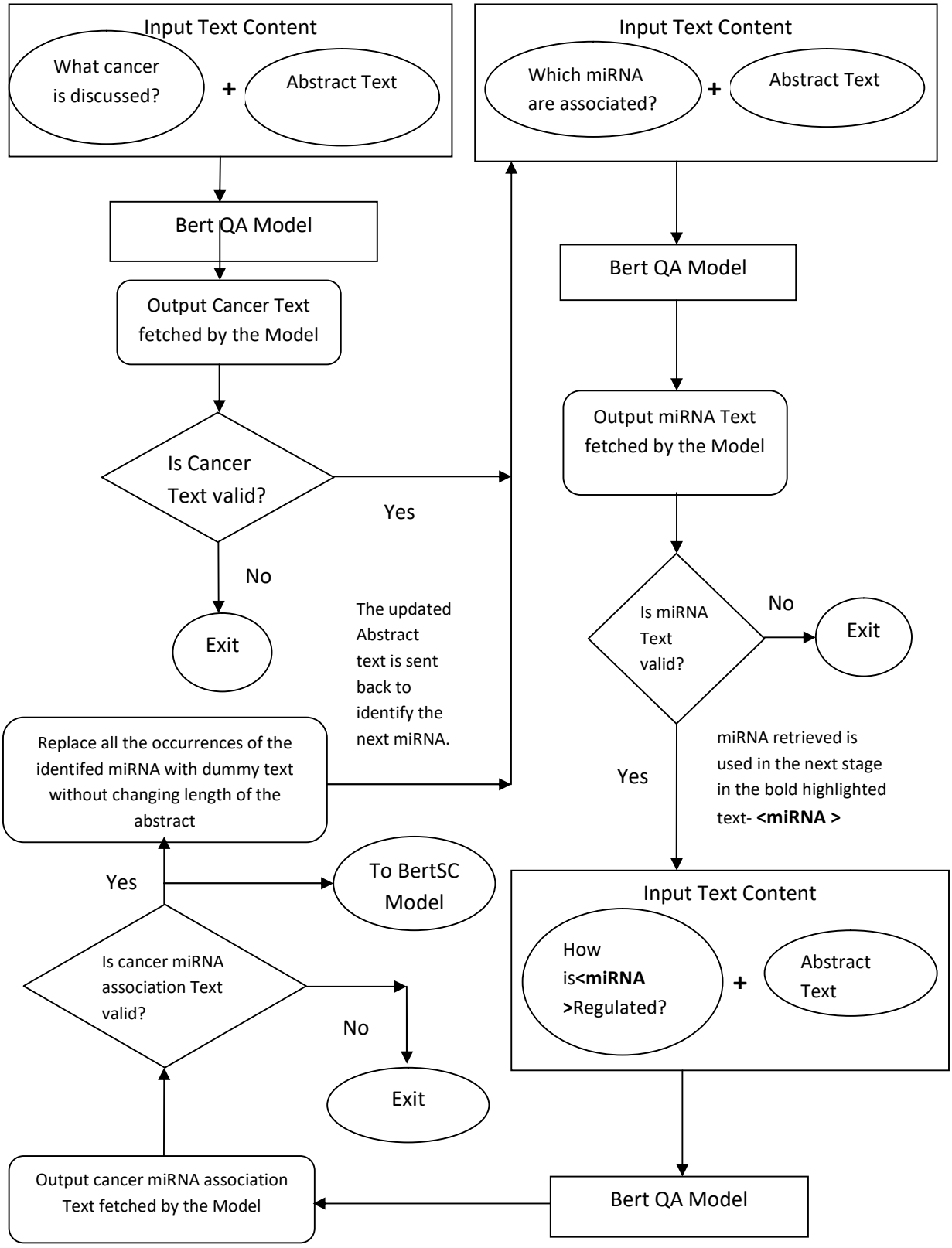
c) Train the bert model using the created training dataset.

5.2 Design the Question Answering training strategy

We will implement a question answering Model based on BERT to extract the required information from the abstract text, using the training strategy described in Figure No 3- Bert Question and Answering Training Strategy:

1. The input text content is first concatenated with the question “What cancer is discussed?”. The answer to this question is specified as the cancer text mentioned in the input text content. If there is no valid cancer text available in the text, we will skip the remaining steps and move to the next input text content.
2. If a valid cancer text is available in the given input text content, the answer to this question is specified as the first miRNA name mentioned in the input text content.
3. We will use the identified miRNA in place of highlighted text in bold in the following question -“How is **<miRNA>** regulated?”. The abstract text is then concatenated with this question. The answer to this question is specified as the cancer regulation behavior mentioned in the input text contentfor the identified miRNA.

Figure No 3- Bert Question and Answering Training Strategy



4. Now Replace all the occurrences of the retrieved miRNA in the given input text content with the dummy text without changing the length of the input text content. We achieve this by creating a dummy text exactly in same length of the identified miRNA text length by concatenating the text "miRNA" followed by the required number of alphabet 's' to match the same length of the identified miRNA name.

   For example, if the identified miRNA text is 'mir-7-5p' which is of 7 characters in length, we will replace all the occurrences of 'mir-7-5p' in the abstract text with 'miRNA' followed by 2 's' (.i.e. miRNAss).
5. Now we will repeat the Steps 2-4 to find the next unique miRNA name mentioned in the given input text content. This will be repeated until there is no valid miRNA name is available in the updated input text content.

5.3 Creating Training Dataset for Q&A Model:

We created training dataset in the SQUAD format in order to train the Bert QA model as mentioned in the below steps. We analyzed all 5300+ abstract texts (from training dataset) and determined the following 4 information –

1. type of cancer
2. name of miRNA
3. actual text in the abstract which describes the cancer-miRNA association
4. cancer-miRNA association- UP or DOWN

We then consolidated the above details for all 4679 research papers in the 5 csv files mentioned earlier which were created for the preparation of training dataset. In the abstract dump, we combined the topic of the research paper and the abstract content into one text which is used as the input text content for the model. We created a python script in which will read the 5 csv files and generate the training dataset in SQUAD format mentioned below.

Below is the template SQUAD format of the training dataset usedused for Bert QA model. Please refer to for sample content of the training dataset in SQUAD format.

SQUAD format template:

```
{
   "version": "v2.0",
   "data": [
      {
         "title": "your_title",
         "paragraphs": [
            {
               "qas": [
                  {
            "question": "Question to ask?",
                     "id": "uniqueId",
                     "is_impossible": "",
                   "answers": [{"text": "your answer", "answer_start": start position of answer in paragraph}]
                  },

               ],
               "context": "astract content paragraph"
            }
         ]
      }
   ]
}
```

5.4 Train the BertQAModel using the created training dataset:

Out of the available flavors of the Bert model offered by Google, we selected the Bert base-uncased model to be used for this Thesis work. This Bert base-uncased model need less computation resources due to less parameters and observed to be working suitable for this thesis work. Choosing a Bert model with higher parameters may need more computation resources and training time but it is possible to improve on prediction accuracy of the model. Also the bertuncased model chosen for this work will analyze the input text contentin a case insensitive fashion.

First, set up the below packages in the Google colab environment

1. Python V3.6 comes by default in Google colab
2. Download the source code of the Huggingface transformers (V3.5.1) into the Google colab machine
3. Install the Huggingface transformers (V3.5.1) using pip python utility
4. Connect to the Google cloud storage console using the required authentications.

Once we have the above setup ready, we followed the below steps in order to train the BertQA model:

1. Upload the created SQUAD training dataset onto the google colab machine.
2. Create the target output folder where the output bert model should be saved after training in the Google colab.
3. Now change the current directory to the path where you have stored the source code of the Huggingface transformers package. Then change the current directory to the target folder which contains the runsquad.py file using the below command:

   Command:

   cd <local path where source code of the Huggingface transformers is saved)/transformers/examples/question-answering
4. Now train the BertQA model by executing the below command with the given hyper parameters.

   Command:

   python run_squad.py --model_typebert --model_name_or_path<local folder path to Input Bert Model> --do_train --do_eval --do_lower_case --train_file<local folder path to the train-v2.json> --predict_file<local folder path to the dev-v2.json> --output_dir<local folder path where the output model should be saved>

   Hyper Parameter Details:

   Below are the details of the hyperparameters used by the BertQA model for this thesis.

   1. per_gpu_train_batch_size = 15
   2. num_train_epochs = 50
   3. max_seq_length = 512

4. doc_stride = 128
5. learning_rate = 0.000005
6. save_steps = 2000000

5. After training is completed, copy the model saved in the target output folder into the desired folder path in the google cloud storage.
6. For the next set of training, we will download the model saved in the google cloud storage from previous training onto the Google colab machine. We use this downloaded Bert model as the input Bert model for the next set of training. We achieve this by specifying the location path of the downloaded Bert model as the value for the parameter in the command used for training the BertQA model.

Chapter 6 - Setting up and finetuning the Bert- Sentence Classification Model

6.1 Finetuning the First Bert Model for Sentence Classification task

Below are the list of tasks performed in order to create the required Bert Sentence Classification model specified the architecture.

a) Design the sentence classification training strategy

b) Create the training dataset for sentence classification Model

c) Train the Bert SC Model using the created training dataset.

6.2 Design the Sentence Classification training strategy :

We implemented a sentence classification model based on BERT to classify the miRNA regulation text output retrieved by the BertQA model from the abstract text as either ‘UP’ or ‘DOWN’ regulated.Below is the training strategy described in Figure No 4- Bert Sentence Classification Training Strategy

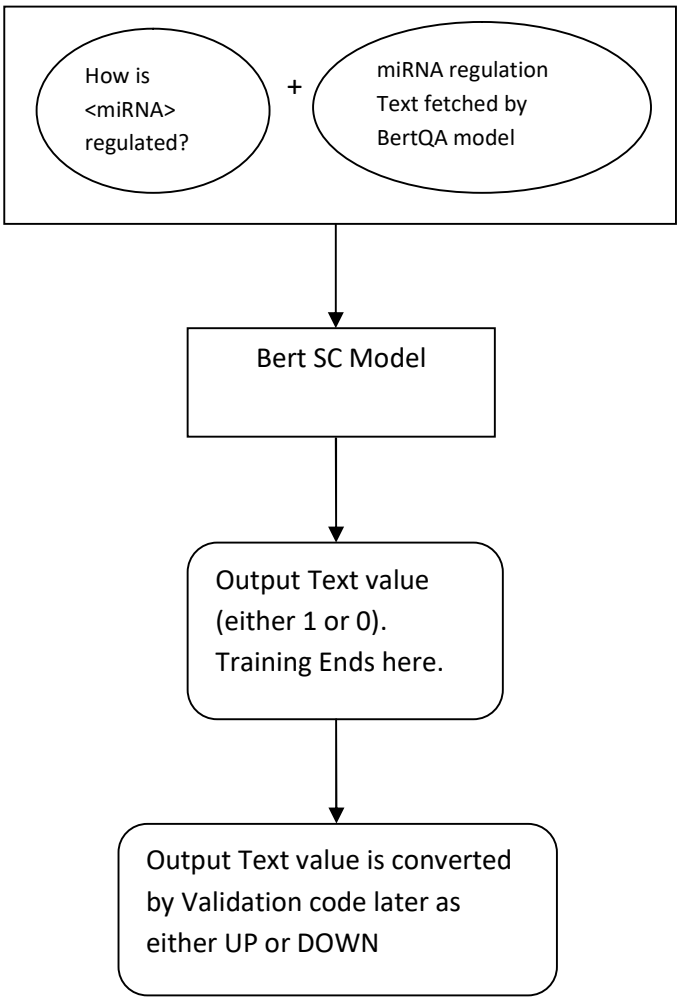
The miRNA regulation textis first concatenated with the question “How is <miRNA> regulated?”. Please note that the actual miRNA identifed by the BertQA model corresponding to the miRNA regulation text is populated in place of the term “<miRNA>” in the above question. The answer to this question is specified as the either ‘1’ or ‘0’.

If the miRNA regulation text is classified as ‘UP’ regulated, we specify the specify the answer to the question as ‘1’.

If the miRNA regulation textis classified as ‘DOWN’ regulated, we specify the specify the answer to the question as ‘0’.

The output text of the BertSC model is always either ‘1’ or ‘0’. This value is converted to ‘UP’ or ‘DOWN’ by the validation code which is explained later in this document.

Figure No 4- Bert Sentence Classification Training Strategy



6.3 Creating Training Dataset for Sentence Classification Model:

We had earlier analyzed all 5300+ abstract texts (from training dataset) and created dataset with 5 csv files. We used the information from the above two csv files and created the dataset required for training the BertSC model in the .tsv (tab separated) format.

1. miRNARegulationDump.csv → This csv file contains details of title of the research paper and regulation type (UP or DOWN regulated) of each miRNA studied in the research paper.
2. RegulationTextFinal.csv → This csv file contains details of the actual text in the input text content describing if the cancer-miRNA association is UP or DOWN regulated

Below is the sample content of the .tsv file prepared for training the Bert SC model

Table-2 (Tsv file content for training Bert SC model)

miRNA Regulation Text from Abstract	Classification Type
How is hsa-mir-375 regulated? Our data showed that miR-146, miR-375, and Let-7 were down-regulated and miR-19 and miR-21 were up-regulated in GC patients with H. pylori infection.	0
How is hsa-mir-19 regulated? Our data showed that miR-146, miR-375, and Let-7 were down-regulated and miR-19 and miR-21 were up-regulated in GC patients with H. pylori infection.	1

6.4 Train the Bert SC Model using the created training dataset:

We again selected the Bert base-uncased model as the base model to create the required Bert SC model. We completed the below same setup in the Google colab environment which we did earlier for training the BertQA model.

1. Python V3.6 comes by default in Google colab
2. Download the source code of the Huggingface transformers (V3.5.1) into the Google colab machine
3. Install the Huggingface transformers (V3.5.1) using pip python utility
4. Connect to the Google cloud storage console using the required authentications.

Once we have the above setup ready, we followed the below steps in order to train the BertSC model:

1. Upload the created .tsv format training dataset onto the google colab machine.
2. Created the target output folder where the output Bert model should be saved after training in the Google colab.

3. Now change the current directory to the path where you have stored the source code of the Huggingface transformers package. Then change the current directory to the target folder which contains the runglue.py file using the below command:

   Command:

   cd <local path where source code of the Huggingface transformers is saved>/transformers/examples/question-answering

4. Now train the BertSC model by executing the below command with the given hyper parameters.

   Command:

   python /content/transformers/examples/text-classification/run_glue.py --task_name SST-2 --model_name_or_path<local folder path to Input Bert Model> --do_train --do_eval --data_dir<local folder path to the input training dataset> --output_dir<local folder path where the output model should be saved>

   Hyper Parameter Details:

   Below are the details of the hyperparameters used by the BertSC model for this thesis.

   a) max_seq_length = 128
   b) per_device_train_batch_size = 32
   c) learning_rate = 2e-5
   d) num_train_epochs = 100

5. After training is completed, Copy the model saved in the target output folder into the desired folder path in the google cloud storage.
6. For the next set of training, we will download the model saved in the google cloud storage from previous training onto the Google colab machine. We use this downloaded bertmodel as the input Bert model for the next set of training. We achieve this by specifying the location path of the downloaded Bert model as the value for the parameter in the command used for training the BertSC model.

Chapter 7 - Validation of the Model and Results Observed

7.1 Validation of the Trained Model

The below tasks are performed in order tovalidate the Bert machine learning model that we had designed and trained earlier

1. Develop an automated solution to validate the final actual output values of the machine learning against the expected results.
2. Create the separate set of test dataset to evaluate the machine learning model.
3. Validate the created model against the test dataset using the developed automated validation solution.
4. Develop an automated solution which will generate output files with cancer-miRNA relationships predicted by the model for the provided list of abstract dumps, without any expected results to compare with.

7.2 Develop an automated solution to validate the final actual output values of the machine learning against the expected results.

As per the training strategy discussed earlier, we train BertQA and BertSC models independently using the training dataset. In order to validate the prediction accuracy of the fully proposed machine learning model, we created a python script which will actually implement the proposed Machine Learning model using the trained BertQA and BertSC models. Please find below the details of the working of this script.

1. The script reads the below 4 csv files (out of the 5 csv files created earlier) one abstract text at a time and feeds the same into the BertQA model. (listed in APPENDIX A)
    I. AbstractDump.csv → This csv file contains details of title of the research paper and input text content of the research paper.
    II. cancerDump.csv → This csv file contains details of title of the research paper and type of cancer studied in the research paper
    III. miRNADump.csv → This csv file contains details of title of the research paper and all the miRNAs studied in the research paper.

IV. miRNARegulationDump.csv → This csv file contains details of title of the research paper and regularion type (UP or DOWN regulated) of each miRNA studied in the research paper.

2. The script first fetches the details of type of cancer questioning the BertQAmodel on cancer type for the given input text content.
3. Using the BertQA model, the script then fetches the details of the unique miRNA names (one at a time) and then followed by corresponding cancer-miRNA association details for the given input text content.
4. Each valid cancer-miRNA association text retrieved by the BertQA model is provided as input to the BertSC model and the final cancer-MiRNA association is determined as either UP or DOWN regulated.
5. Now the script validates if the final values of the 3 Information types predicted by the proposed machine model matches with the expected values for the same as mentioned in the originally provided input csv files below
   I. cancerDump.csv -The expected correct value of cancer type is available in this csv file
   II. miRNADump.csv - The expected correct value of the list of miRNA names are available in this csv file
   III. miRNARegulationDump.csv - The expected correct value of miRNARegulation type (UP or DOWN) for each unique miRNA is available in this csv file.

6. The script creates a file-Validation Errors.csv to log details of any mismatch between the expected and actual values of any information type for the given abstract text.

7. The script also creates the file –Outputmetrics.csv (listed in APPENDIX C) with the following details:
   I. total number of expected correct answers for each input text content provided in the input csv files
   II. total number of actual correct answers predicted by the model for each input text content provided in the input csv files

8. The script also creates the below 3 csv files (listed in APPENDIX C) with the actual output values of the 3 information types predicted by the proposed Machine learning model for each input text content given in the input csv files.
   I. Cancer_output.csv
   II. Mirna_output.csv
   III. Mirna_regulation_output.csv

7.3 Create the separate set of test dataset to evaluate the machine learning model

Just like we created the csv files for training dataset, we also created the below 4 csv files (listed in APPENDIX A) using abstract content of 1145 research papers which was used as the test dataset.

1. AbstractDump.csv
2. cancerDump.csv
3. miRNADump.csv
4. miRNARegulationDump.csv

7.4 Validate the created model against the test dataset using the developed automated validation solution.

We validated the created machine learning model using the automated validation python script and the 4 csv files created using the test dataset. As mentioned earlier, the test dataset consisted of 1145 abstract texts. We observed the proposed machine learning model was able to predict the correct values of the 3 information categories which are required to determine the cancer miRNA association with an overall accuracy of 90%.The Table -3 (Validation Results of the Machine Learning Model against Test Dataset)shows the details of the results observed from the validation of the model against test dataset.

Table -3 (Validation Results of the Machine Learning Model against Test Dataset)

Metrics from validation results	Value
Total number of cancer available in the test dataset	1145
Number of correct cancer predictions	1131
Total number of miRNA available in the test dataset	1213
Number of correct miRNA predictions	1195
Total number of miRNA regulation available in the test dataset	1213
Number of correct miRNA regulation predictions	1107
Total number of predictions made by the model in the test dataset	3571
Total number of correct Predictions	3433
Percentage of correct cancer predictions	98.78%
Percentage of correct miRNA predictions	98.52%
Percentage of correct miRNA regulation predictions	91.26%
Percentage of total correct predictions	96.14%
Total number of cancer-mirna relation available in the test dataset	1145
Total number of cancer-mirna relations predicted correctly (.i.e. all the 3 types of predictions-cancer, miRNAs, miRNA regulations are correct for a topic)	1035
Percentage of correct cancer-mirna relation predictions in the test dataset	**90.39%**

We also observed that the model predicted all 3 Information categories correctly for 90.39% of the abstract texts.

7.5 Develop an automated solution which will generate output files consisting of all cancer-miRNA relationships predicted by the model without any validations

The functionality of validation script created earlier (in section 7.2) is to compare if the cancer-miRNA relationships predicted by the model for the list of abstracts provided matches the expected results. This script will be useful if we have the expected results are already available (from other methods of text mining) and we plan to use the model only to review if the expected results are correct. We developed another automation script which can be used in case the expected results are not available and the requirement for the model is actually to extract the cancer-miRNA relations from the given list of abstract texts. To summarize, we have 2 types of scripts (both are listed in **Error! Reference source not found.**)) which can be used with the model-

1. One script is used to validate if the cancer-miRNA relationships provided in the input csv files are correct based on the predictions made by the model.

2. Other script is used to extract cancer-miRNA relationships from the given list of abstract text content but has no expected results to compare with.

Chapter 8- Conclusion

8.1 Highlights of the current work and scope for future improvements

In this thesis, we have successfully designed and implemented a machine learning model using Bert framework to extract the information of cancer-miRNA association from the given input text content (title + abstract text). We have excluded analyzing any input text content of size greater than 512 tokens (as per the bert vocabulary) due to limitation of Bert model in handling texts larger than this specified size. The machine learning model created in this thesis is observed to function with prediction accuracy of 90.39% (against test dataset). It can serve as a quick and effective means of extracting cancer-miRNA association details from the research papers. As the model is implemented and trained directly from the abstract text, it will learn to accomodate any future texts as when continuously train the model using future literatures and it does not require any manual analysis of the text to add new rules. Also this machine learning model based approach can be used to review if the results retrieved by the existing mircancer text mining implementation are correct. The cancer-miRNA relationships identified by the model can be compared with that of the results retrieved by the existing mircancer text mining implementation and the matching results can be considered as True Positives. Then we need to review only the remaining literatures for which the results do not match between the two solutions. The model developed by us is designed to handle input literatures which study on a single type of cancer. In case there is more than one type of cancer studied in the research paper, the model will retrieve only the first type of cancer mentioned in the input text content.We can extend this model to handle multiple types of cancer mentioned in the given abstract as part of future scope of work on this topic. The work done in this thesis to generate the training and test dataset could be of good use for future enhancements in this area.

REFERENCES

1. miRCancer: a microRNA–cancer association database constructed by text mining on literature. (by Boya Xie, Qin Ding, Hongjin Han, Di Wu)
2. Chen,C. et al. (2004) MicroRNAs modulate hematopoietic lineage differentiation. Science, 303, 83–86.
3. Dweep,H. et al. (2011, May) miRWalk–database: prediction of possible miRNA binding sites by ‘‘walking’’ the genes of three genomes. J. Biomed. Inform., 44, 839–847.
4. Fritz,A. et al. (2000) International Classificaiton of Diseases for Oncology, 3rd edn. World Health Organization, Geneva.
5. Hsu,S. et al. (2011) miRTarBase: a database curates experimentally validated microRNA-target interactions. Nucleic Acids Res., 39, D163–D169.
6. Hwang,H.-W. and Mendell,J.T. (2006) MicroRNAs in cell proliferation, cell death, and tumorigenesis. Br. J. Cancer, 94, 776–780.
7. Iorio,M.V. et al. (2005) MicroRNA gene expression deregulation in human breast cancer. Cancer Res., 65, 7065–7070.
8. Jiang,Q. et al. (2009) miR2Disease: a manually curated database for microRNA deregulation in human disease. Nucleic Acids Res., 37, D98–D104.
9. Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation. (by Jacob O'Brien, Heyam Hayder, Yara Zayed, Chun Peng)
10. http://mircancer.ecu.edu/
11. https://github.com/google-research/bert
12. https://huggingface.co/transformers/
13. https://www.frontiersin.org/articles/10.3389/fendo.2018.00402/full
14. https://arxiv.org/abs/1908.08962
15. https://arxiv.org/abs/1810.04805
16. https://clinicalepigeneticsjournal.biomedcentral.com/articles/10.1186/s13148-018-0587-8/figures/1
17. https://pubmed.gov

18. https://hackernoon.com/nlp-tutorial-creating-question-answering-system-using-bert-squad-on-colab-tpu-1utp3352
19. https://www.analyticsvidhya.com/blog/2020/07/transfer-learning-for-nlp-fine-tuning-bert-for-text-classification/
20. https://towardsdatascience.com/getting-started-with-google-colab-f2fff97f594c
21. https://www.tutorialspoint.com/vba/index.htm
22. https://www.tutorialspoint.com/java/index.htm
23. https://www.tutorialspoint.com/python/index.htm
24. https://www.w3schools.com/jsref/jsref_obj_regexp.asp

APPENDIX A

1. Refer the below table for the sample content of AbstractDump.csv:

Topic	Abstract
MicroRNA let-7a suppresses breast cancer cell migration and invasion through downregulation of C-C chemokine receptor type 7.	C-C chemokine receptor type 7 (CCR7) plays an important role in chemotactic and metastatic responses in various cancers, including breast cancer. In the present study, the authors demonstrated that microRNA (miRNA) let-7a downregulates CCR7 expression and directly influences the migration and invasion of breast cancer cells.The expression of CCR7, its ligand CCL21, and let-7a was detected in breast cancer cell lines and in breast cancer patient tissues. Synthetic let-7a and an inhibitor of let-7a were transfected into MDA-MB-231 and MCF-7 breast cancer cells, respectively, and cell proliferation, cell migration, and invasion assays were performed. To confirm the fact that 3'UTR of CCR7 is a direct target of let-7a, a luciferase assay for the reporter gene expressing the let-7a binding sites of CCR7 3'UTR was used. An in vivo invasion animal model system using transparent zebrafish embryos was also established to determine the let-7a effect on breast cancer cell invasion.First, a higher expression of both CCR7 and CCL21 in malignant tissues than in their normal counterparts from breast cancer patients was observed. In addition, a reverse correlation in the expression of CCR7 and let-7a in breast cancer cell lines and breast cancer patient tissues was detected. Synthetic let-7a decreased breast cancer cell proliferation, migration, and invasion, as well as CCR7 protein expression in MDA-MB-231 cells. The let-7a inhibitor reversed the let-7a effects on the MCF-7 cells. The 3'UTR of CCR7 was confirmed as a direct target of let-7a by using the luciferase assay for the reporter gene expressing let-7a CCR7 3'UTR binding sites. Notably, when analyzing in vivo invasion, MDA-MB 231 cells after synthetic let-7a transfection were unable to invade the vessels in zebrafish embryos.The results from the present study suggest that targeting of CCL21-CCR7 signaling is a valid approach for breast cancer therapy and that let-7a directly binds to the 3'UTR of CCR7 and blocks its protein expression, thereby suppressing migration and invasion of human breast cancer cells. Furthermore, the present study underscores the therapeutic potential of let-7a as an antitumor and antimetastatic manager in breast cancer patients.

2. Refer the below table for the sample content of CancerDump.csv below:

Topic	Cancer
MicroRNA let-7a suppresses breast cancer cell migration and invasion through downregulation of C-C chemokine receptor type 7.	breast cancer

3. Refer the below table for the sample content of miRNADump.csv:

Topic	miRNA
MicroRNA let-7a suppresses breast cancer cell migration and invasion through downregulation of C-C chemokine receptor type 7.	hsa-let-7a

4. Refer the below table for the sample content of miRNA_RegulationDump.csv :

Topic	miRNARegulation
MicroRNA let-7a suppresses breast cancer cell migration and invasion through downregulation of C-C chemokine receptor type 7.	down

5. Refer the below table for the sample content of RNA_RegulationFinal.csv:

Topic	miRNARegulationText
MicroRNA let-7a suppresses breast cancer cell migration and invasion through downregulation of C-C chemokine receptor type 7.	Furthermore, the present study underscores the therapeutic potential of let-7a as an antitumor and antimetastatic manager in breast cancer patients.

APPENDIX B

Refer the below JSON content for the sample training file content (in squad format) used to train the

BertQA model


```
{
  "version":"v2.0",
  "data":[ {
      "title":"",
      "paragraphs":[ {
          "qas":[ {
              "question":"What cancer is discussed?",
              "id":"1",
              "is_impossible":false,
              "answers":[ {
                  "text":"breast cancer",
                  "answer_start":131
                } ]
            } ],
          "context":"C-C chemokine receptor type 7 (CCR7) plays an important role in chemotactic and metastatic responses in various cancers, including breast cancer. In the present study, the authors demonstrated that microRNA (miRNA) let-7a downregulates CCR7 expression and directly influences the migration and invasion of breast cancer cells.The expression of CCR7, its ligand CCL21, and let-7a was detected in breast cancer cell lines and in breast cancer patient tissues. Synthetic let-7a and an inhibitor of let-7a were transfected into MDA-MB-231 and MCF-7 breast cancer cells, respectively, and cell proliferation, cell migration, and invasion assays were performed. To confirm the fact that 3'UTR of CCR7 is a direct target of let-7a, a luciferase assay for the reporter gene expressing the let-7a binding sites of CCR7 3'UTR was used. An in vivo invasion animal model system using transparent zebrafish embryos was also established to determine the let-7a effect on breast cancer cell invasion.First, a higher expression of both CCR7 and CCL21 in malignant tissues than in their normal counterparts from breast cancer patients was observed. In addition, a reverse correlation in the expression of CCR7 and let-7a in breast cancer cell lines and breast cancer patient tissues was detected. Synthetic let-7a decreased breast cancer cell proliferation, migration, and invasion, as well as CCR7 protein expression in MDA-MB-231 cells. The let-7a inhibitor reversed the let-7a effects on the MCF-7 cells. The 3'UTR of CCR7 was confirmed as a direct target of let-7a by using the luciferase assay for the reporter gene expressing let-7a CCR7 3'UTR binding sites. Notably, when analyzing in vivo invasion, MDA-MB 231 cells after synthetic let-7a transfection were unable to invade the vessels in zebrafish embryos.The results from the present study suggest that targeting of CCL21-CCR7 signaling is a valid approach for breast cancer therapy and that let-7a directly binds to the 3'UTR of CCR7 and blocks its protein expression, thereby suppressing migration and invasion of human breast cancer cells. Furthermore, the present study underscores the therapeutic potential of let-7a as an antitumor and antimetastatic manager in breast cancer patients."
        } ]
    } ]
}
```

APPENDIX C

1. Refer the below table for the sample content of output.csv below:

Topic	Total Count	Total Right	miRNA Count	miRNA Right	Reg Count	Reg Right
ADAM9 enhances CDCP1 protein expression by suppressing miR-218 for lung tumor metastasis.	3	2	1	1	1	1

2. Refer the below table for the sample content of ValidationErrors.csv below:

Topic	Field	Expected Answer	Actual Answer	miRNA	Comments
ADAM9 enhances CDCP1 protein expression by suppressing miR-218 for lung tumor metastasis.	Cancer	lung cancer	lung tumor	NA	Actual Cancer does not Match with Expected

3. Refer the below table for the sample content of cancer_Output.csv below:

Topic	Abstract	Cancer
ADAM9 enhances CDCP1 protein expression by suppressing miR-218 for lung tumor metastasis.	ADAM9 enhances CDCP1 protein expression by suppressing miR-218 for lung tumor metastasis.Metastasis is the leading cause of death in cancer patients due to the difficulty of controlling this complex process. MicroRNAs (miRNA), endogenous noncoding short RNAs with important biological and pathological functions, may play a regulatory role during cancer metastasis, but this role has yet to be fully defined. We previously demonstrated that ADAM9 enhanced the expression of the pro-migratory protein CDCP1 to promote lung metastasis; however, the regulatory process remains unknown. Here we demonstrate that endogenous miR-218, which is abundant in normal lung tissue but suppressed in lung tumors, is regulated during the process of ADAM9-mediated CDCP1 expression. Suppression of miR-218 was associated with high migration ability in lung cancer cells. Direct interaction between miR-218 and the 3'-UTR of CDCP1 mRNAs was detected in luciferase-based transcription reporter assays. CDCP1 protein levels decreased as expression levels of miR-218 increased, and increased in cells treated with miR-218 antagomirs. Induction of miR-218 inhibited tumor cell mobility, anchorage-free survival, and tumor-initiating cell formation in vitro and delayed tumor metastases in mice. Our findings revealed an integrative tumor suppressor function of miR-218 in lung carcinogenesis and metastasis.	lung tumor

4. Refer the below table for the sample content of miRNA_Output.csv below:

ADAM9 enhances CDCP1 protein expression by suppressing miR-218	ADAM9 enhances CDCP1 protein expression by suppressing miR-218 for lung tumo metastasis.Metastasis is the leading cause of death in cancer patients due to the difficulty of controlling this complex process. MicroRNAs (miRNA), endogenous noncoding short RNAs with important biological and pathological functions, may play regulatory role during cancer metastasis, but this role has yet to be fully defined. We previously demonstrated that ADAM9 enhanced the expression of the pro-migratory protein CDCP1 to promote lung metastasis; however, the regulatory process remains unknown. Here we demonstrate that endogenous miR-218, which is abundant in	mir-218

for lung tumor metastasis.	normal lung tissue but suppressed in lung tumors, is regulated during the process of ADAM9-mediated CDCP1 expression. Suppression of miR-218 was associated with high migration ability in lung cancer cells. Direct interaction between miR-218 and the 3'-UTR of CDCP1 mRNAs was detected in luciferase-based transcription reporter assays. CDCP1 protein levels decreased as expression levels of miR-218 increased, and increased in cells treated with miR-218 antagomirs. Induction of miR-218 inhibite tumor cell mobility, anchorage-free survival, and tumor-initiating cell formation in vitro and delayed tumor metastases in mice. Our findings revealed an integrative tumor suppressor function of miR-218 in lung carcinogenesis and metastasis.	

5. Refer the below table for the sample content of miRNA_Output.csv below:

Topic	**Abstract**	**miRNA**	**Actual miRNA Regulation**
ADAM9 enhances CDCP1 protein expression by suppressing miR-218 for lung tumor metastasis.	ADAM9 enhances CDCP1 protein expression by suppressing miR-218 for lung tumor metastasis.Metastasis is the leading cause of death in cancer patients due to the difficulty of controlling this complex process. MicroRNAs (miRNA), endogenous noncoding short RNAs with important biological and pathological functions, may play a regulatory role during cancer metastasis, but this role has yet to be fully defined. We previously demonstrated that ADAM9 enhanced the expression of the pro-migratory protein CDCP1 to promote lung metastasis; however, the regulatory process remains unknown. Here we demonstrate that endogenous miR-218, which is abundant in normal lung tissue but suppressed in lung tumors, is regulated during the process of ADAM9-mediated CDCP1 expression. Suppression of miR-218 was associated with high migration ability in lung cancer cells. Direct interaction between miR-218 and the 3'-UTR of CDCP1 mRNAs was detected in luciferase-based transcription reporter assays. CDCP1 protein levels decreased as expression levels of miR-218 increased, and increased in cells treated with miR-218 antagomirs. Induction of miR-218 inhibited tumor cell mobility, anchorage-free survival, and tumor-initiating cell formation in vitro and delayed tumor metastases in mice. Our findings revealed an integrative tumor suppressor function of miR-218 in lung carcinogenesis and metastasis.	mir-218	our findings revealed an integrative tumor suppressor function of mir-218 in lung carcinogenesis and metastasis .