Thesis Report

Cancer Classification with Deep Learning using Genomics Data

Ahmad Omar Ahsan
ID: **160041001**
Department of CSE, Islamic University of Technology
ahmadomar@iut-dhaka.edu

Md. Azmaeen Bin Ansar
ID: **160041030**
Department of CSE, Islamic University of Technology
azmaeen@iut-dhaka.edu

Minhajul Islam Minhaj
ID: **160041061**
Department of CSE, Islamic University of Technology
minhajulislam@iut-dhaka.edu

Co-supervisor
Tasnim Ahmed
Lecturer, Department of CSE, Islamic University of Technology
tasnimahmed@iut-dhaka.edu

Supervisor
Tareque Mohmud Chowdhury
Assitant Professor, Department of CSE, Islamic University of Technology
tareque@iut-dhaka.edu

Declaration of Authorship

This is to certify that the work presented in this thesis is the outcome of the analysis and simulations carried out by **Ahmad Omar Ahsan**,**Md. Azmaeen Bin Ansar** and **Minhajul Islam Minhaj** under the supervision of **Tareque Mohmud Chowdhury**, Assistant Professor at Department of Computer Science and Engineering (CSE), Islamic University of Technology (IUT), Dhaka, Bangladesh and co-supervision of **Tasnim Ahmed**, Lecturer at Department of Computer Science and Engineering(CSE), Islamic University of Technology(IUT).

It is also declared that neither of this thesis nor any part of this thesis has been submitted anywhere else for any degree or diploma. Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references is given.

Ahmad Omar Ahsan
StudentID: 160041001

Md. Azmaeen Bin Ansar
Student ID: 160041030

Minhajul Islam Minhaj
Student ID:160041061

Tareque Mohmud
Choudhury ,
Assistant Professor,
Department of CSE,
Islamic University of
Technology.(IUT)

Tasnim Ahmed ,
Lecturer,
Department of CSE,
Islamic University of
Technology.(IUT)

Acknowledgement

At At the outset, we express utmost gratitude to Almighty Allah for His blessings which allowed us to shape this research into reality and give it form.

This thesis owes its existence to a lot of people for their support, encouragement, and guidance. We would like to express our gratitude towards them.

We are very grateful to our supervisor **Tareque Mohmud Chowdhury** , Assistant Professor, Department of Computer Science and Engineering, Islamic University of Technology (IUT), for his supervision, knowledge and support, which has been invaluable for us.This research work would not see the light of success without his supervision, strategy, and line of action.

We are grateful to **Tasnim Ahmed**, Lecturer, Department of Computer Science. He has been a constant source of enthusiasm and encouragement which has allowed us to turn this endeavor into a success story.

Finally, we seize this opportunity to express our profound gratitude to our beloved parents for their love and continuous support both spiritually and mentally.

Abstract

Deep learning has been monumental in Computer Vision, Natural Language Processing, Machine Translation task and so on. In bioinformatics, Deep learning is playing an important role in drug discover and protein structure prediction. In cancer diagnosis, thanks to advances in Computer Vision Deep Learning models are able to accurately classify cancer. However, not much work has been done in the field of Cancer diagnosis with genomic data. Several authors attempted to use genomic data using machine learning, however it was restricted to single cancer subtypes. In this thesis, we explored classification of all types of cancer using miRNA genome data by creating new model architectures.

We are proposing two new architectures a basic ANN and a novel architecture based on ResNet called CResNet. We have trained 4 different kinds of model. LSTM, Artificial Neural Network, CResNet (Variant of ResNet Architecture) and Ensemble models using model averaging. Our models have achieved MCC (Mathew's correlation coefficient) value of 0.8596,0.9625,0.9745 and 0.9439 which is greater than the SOTA model's MCC model demonstrating that our architecture performed better than the current architecture.

Contents

List of Tables

List of Figures

1 Introduction

The domain of Computer Science most relevant to our research is Bioinformatics. Bioinformatics is the study of the process of extraction of biological data, converting it into meaningful forms and finally the analysis of said data. The specific part of Bioinformatics that we have focused on is cancer subtype diagnosis. The data that will be used for this diagnosis is miRNA expression data. Our rudimentary models will employ techniques such as LSTM (Long Short Term Memory) models and NCA (Neighborhood Component Analysis), but this can be subject to future change and we also are using ANN. The objective of our project is to build an efficient ML/DL model to accurately detect the various cancer types and subtypes based on the miRNA expression data.

In this research we are trying to incorporate deep learning and bioinformatics in such a way so that we can give cancer diagnosis based on gene expression of the cancer. Here we are implementing LSTM which is a sequence model and ANN which are artificial neural network on micro RNA sample to classify all cancer types and subtypes. Our approach will use all the available features from the micro RNA. Then using those features we will perform classification. After classification we will evaluate the performance of our model using confusion matrix and Mathew's coefficient. We are also implementing CResNet (a variant of ResNet) and Ensemble model averaging on miRNA expression quantification data for classification of 8 cancer types and subtypes

Our research will bring new avenues and ways to use deep learning in classification for cancer. Existing classification methods of cancer heavily relies on Convolutional Neural Network which deals with images. Our approach will present a way to preprocess data, analyze micro RNA reads and use that data to perform classifications using sequential models.

In short, our research goal is to use deep learning models to evaluate the performance of classification on all types of cancer.

2 Motivation and Problem Statement

Cancer is the second leading cause of death globally, and is responsible for an estimated 9.6 million deaths in 2018. Better understanding of etiological and biological nature of cancer will equip us with better methods of preventive, diagnostic and therapeutic tools to help reduce the burden of this disease.

Early diagnosis of cancer is extremely important when it comes to treatment. When cancer is found at an early stage before it has spread, the 5-year relative survival rate is about 80%. But only cervical cancer and breast cancer can be detected early. When cancer has spread outside the affected area, survival rates are lower. Therefore, current screening and detection technologies being more competent than ever before are still very limited and there is significant room for improvement. Adding to this, Cancer is a heterogeneous disease and many cancer types do not represent a single entity, but are composed of biologically and clinically diverse subtypes. Different treatment strategies exist for different subtypes and therefore it is important to ascertain the subtype as well. Machine Learning/Deep Learning algorithms are a burgeoning entity in various fields of health sciences. An instance would be the extensive use of computer vision technologies in determining various malignant cell growth from imaging with greater specificity than a specialist. Machine learning techniques are also used in many bioinformatics research to mine new knowledge on existing sequence data. Our research employs a similar approach. The problem that our research aims to tackle is the accurate diagnosis of cancer types and subtypes.

LSTM is a recurrent neural network algorithm that is capable of learning order dependence on sequence prediction problems. Our goal as researchers is to discover the potential of this LSTM algorithm in cancer subtype detection by running it on microRNA expression data. microRNA expression levels in various cancers serve as informative biomarkers and can also be obtained from patients in non invasive methods. Not all microRNA have equal weight as cancer biomarkers, therefore it is important to determine the most impactful microRNA in diagnosis. Finally our problem statement stands as: Finding the most impactful microRNA in terms of acting as biomarkers and then using said microRNA data to classify cancer types and subtypes .
ANN stands for artificial neural network. The idea is to introduce non linear activation function to capture non linear patterns in miRNA quantification data. It is of great importance that we capture non linear activation pattern the reason being that all types of cancer will be classified. While cancer subtypes might show linear patterns, different cancer subtypes will show non linear pattern. We believe that ANN will be able to capture those pattern and offer greater classification accuracy.

CResNet is a variant of ResNet model architecture. ResNet is a popular model architecture that uses skip connections. It is an important feature for deep neural network architectures. This is because as model's get deeper and deeper some features might get lost. So skip connections retains the information and then adds with the activation calculation performed on the information. This retains some feature which might get lost as the architecture gets deeper and deeper.

We are also building an ensemble of deep neural networks. We are going to use model voting which takes the most common prediction and gives the output. Ensemble techniques are popular in Kaggle a site where Ml competition is hosted.

We propose to build a stage independent deep learning models for early detection and classification of cancer.

3 Related Works

In this section we are going to discuss about related works related to our research, background study as to why microRNAs are the perfect biomarker for Cancer classification.

3.1 The role of microRNAs in Cancer [15]

MicroRNAs are small, 18–24 nucleotide RNAs that regulate the translation and stability of specific target mRNAs.They are responsible for regulating a wide array of biological processes including carcinogenesis. In cancer cells. it has been observed that miRNAs are heavily dysregulated.

They are dyregulated in human cancer through various mechanisms,including amplification or deletion of miRNA genes,abnormal transcriptional control of miRNAs, dysregulated epigenetic changes and defects in the miRNA biogenesis machinery. These dysregulation will make miRNAs function as oncogenes or tumor suppressors under certain conditions.

3.1.1 miRNA biogenesis and Regulation

The miRNA biogenesis begins with transcribing gene into large primary transcript (pri-miRNA), which is 5 capped and 3 polyadenylated in structure. transcription is typically mediated by RNA polymerase3. The pri-miRNA are cleaved by a microprocessor complex, composed of RNA-binding protein into a 85-nucleotide stem-loop structure called precursor miRNA. After transporting from nucleus to cytoplasm by GTP 5 complex, it is then processed by another RNA polymerase into 20-22 nucleotide, miRNA duplex. Then it is unwound and and the mature miRNA is incorporated into a protein complex known as RNA-induced silencing complex (RISC) and guides it to target miRNA. The mature miRNA can be degraded to form RNAs and it can also combine with TLR to trigger downstream signaling pathways.

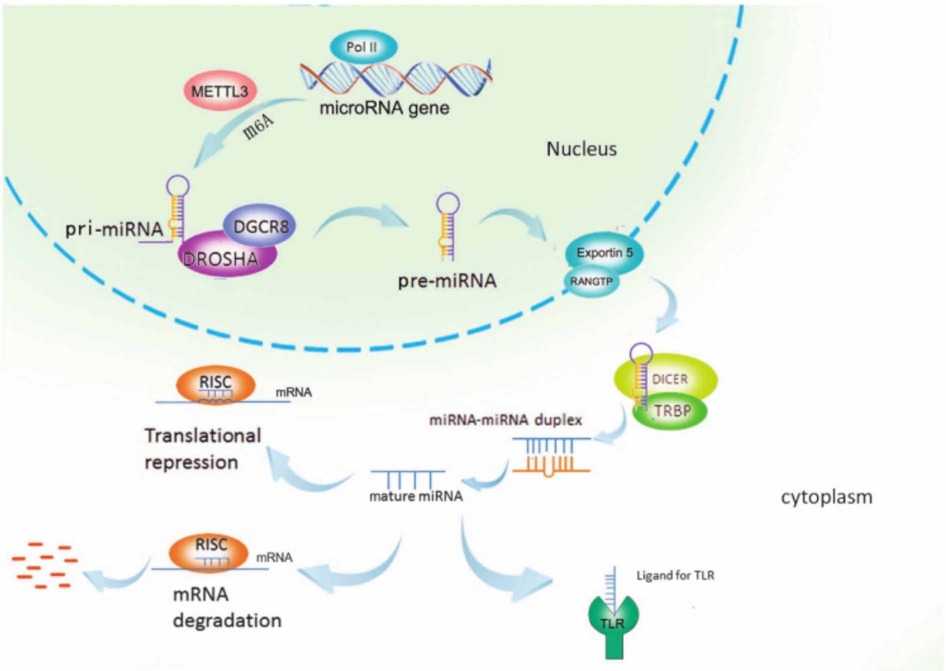


Figure 1: miRNA biogeneis

3.1.2 Mechanisms of miRNA dysregulation in cancer

Alterations in genomic miRNA copy numbers and gene locations are the main reasons behind abnormal miRNA expression in malignant cells compared with normal cells. Earliest discovery of miRNA gene location change is the loss of miR-15a/16-1 cluster gene at chromosome 13q14, which is frequently observed in leukemia patients. In lung cancer, the 5q33 region harboring miR-143 and miR-145 is often deleted,resulting in decreased expression of both miRNAs [1]. Amplification of miR-17-92 cluster gene was observed in B-cell lymphomas[20] and lung cancers [7], and translocation of this cluster gene was also observed in T-cell acute lymphoblastic leukemia, leading to over expression of these miRNAs in these malignancies[12].High frequency of genomic alterations in miRNA loci was confirmed by high-resolution array-based comparative genomic hybridization in 227 specimens from human ovarian cancer, breast cancer and melanoma[23]. Overall these findings suggest that abnormal miRNA expression in malignant cells could arise from amplification or deletion of specific genomic regions encompassing miRNA genes.

3.1.3 Significance of the altered miRNA expression in tumors

It has been proposed by Hanahan and Weinberg [6] that the hallmarks of human cancer comprise of six biological capabilities they are:

- Sustaining proliferative signaling
- Evading growth suppressors
- Resisting cell death
- Enabling replicative immortality
- Activating invasion and metastasis
- Angiogenesis

Evading growth suppressors and sustaining proliferative signaling Cell proliferation is the most important hallmark of cancer and its abnormality is the leading cause of tumorigenesis. In details, cell- cycle progression is controlled by intracellular programs and extracellular signal molecules, to reach the balance between promoting cell proliferation and suppressing it. Cells become cancerous when cell growth or division is out of control. Over the years of studies, it becomes apparent that some miRNAs functionally integrate into multiple critical cell proliferation pathways, and the dysregulation of these miRNAs is responsible for evading growth suppressors and sustaining proliferative signaling in cancer cells.

Evasion of apoptosis is another significant hallmark of tumor progression, which is believed to be regulated by miRNAs. Tumor cells evolve a variety of strategies to limit or circumvent apoptosis. Among them, the loss of p53 tumor suppressor function is most common. The alternative ways to evade apoptosis include upregulation of anti-apoptotic regulators, suppression of proapoptotic factors and inhibition of death pathway induced by extrinsic ligands. The components involved in anti-apoptosis are broadly inhibited or activated by miRNAs.

Metastasis is a complex, multistep and dynamic biological event. Epithelial–mesenchymal transition (EMT) is considered an early and key step in the metastatic cascade, characterized by loss of cell adhesion through repression of E-cadherin and activation of genes associated with motility and invasion. EMT is thought to be regulated by a variety of signaling pathways such as transforming growth factor (TGF)-β, all of which converge on the key transcription factors such as ZEB, SNAIL and TWIST.

Since the discovery of miR-15a and miR-16-1 deletions in chronic lymphocytic leukemia, many laboratories around the world have demonstrated the expression of miRNAs is dysregulated in different tumors. Such dysregulation could be caused by multiple mechanisms, including amplification or deletion of miRNA genes, abnormal transcriptional control of miRNAs, dysregulated epige- netic changes and defects in the miRNA biogenesis machinery. Cancer cells with abnormal miRNA expression evolve the capability to sustain proliferative signaling, evade growth suppressors, resist cell death, activate invasion and metastasis and induce angiogenesis. MiRNA may function as either tumor suppressor or oncogene under certain circumstances. Genome-wide profiling demonstrates that miRNA expression signatures are associated with tumor type, tumor grade and clinical outcomes, so miRNAs could be potential candidates for diagnostic biomarkers, prognostic biomarkers, therapeutic targets or tools.

3.2 The Role of microRNAs in Colorectal Cancer [17]

MicroRNAs are small, 18–24 nucleotide RNAs that regulate the translation and stability of specific target mRNAs.Almost 18 years ago, microRNAs, were implicated in the initiation of chronic lymphocytic leukemia [2]. Since that discovery, microRNAs have been shown to be involved in almost every aspect of cancer biology, as tumor suppressor genes or oncogenes depending on the cellular context in which they are expressed. Evidence supports a role for microRNAs at every stage of CRC initiation, progression and development. Extensive research were aimed at determining if microRNAs can be used as diagnostic biomarkers. In this review paper, the role of microRNA in CRC was mainly discussed in different phases of CRC.

3.2.1 MicroRNA expression is consistently altered in CRC

More than 20 studies were taken when the paper was released. Those studies have examined microRNA expression patterns in CRC and confirmed that microRNAs have consistently and reproducibly altered in CRC [10]. These studies used a variety of techniques ranging from global miRNA expression profiling with deep sequencing[3] or microRNA microarrays[9] to examine the expression of selected microRNAs with quantitative reverse transcriptase polymerase chain reaction (qRT-PCR). The main underlying theme of these studies is that microRNA expression of CRC is distinctly different than non tumor tissues, which is consistent with the hypothesis that aberrant microRNA expression has a role in CRC initiation and development.

In contrast to the original report that microRNA expression levels are globally reduced in cancer,[9], more microRNAs have been found to have elevated expression in CRC compared to those with reduced levels. A review of 23 micro RNA expression studies found that of the 164 microRNAs that are significantly altered in CRC in at least one study, approximately 2/3 of them were elevated and 1/3 that were reduced in tumors[10]. This indicates that microRNAs may have more oncogenic than tumor suppressive functions in CRC. This means that certain microRNAs have important oncogenic functions while others have important tumor suppressor functions and these functions need to be evaluated for each microRNA individually in the context of the specific tissue/tumor type.

MicroRNA expression patterns can also classify tissue types and tumor types and microRNA expression patterns perform at least as well as mRNA expression profiles for this purpose.[9].Therefore, microRNA expression pattern may help classify different phenotypic subgroups of CRC.

3.2.2 MicroRNAs function at early stages in CRC development

Adenomas are benign growths that are frequently a precursor lesion of colon adenocarcinoma. If microRNAs are altered in adenomas, it suggests that microRNAs have a role in the initiation of cancer. This is indeed the case. MicroRNA expression patterns can distinguish normal colonic mucosa, colon adenomas, and colon carcinoma [14]. These expression patterns are consistent with the stepwise, multi-hit model for colon carcinogenesis and support a role for microRNAs in each step. MiR-21 is a good example as it is elevated in adenomas and colon carcinomas [16].Higher expression levels of miR-21 correlate with advance stages of CRC indicating a role for miR-21 in initiation and progression of CRC [16].

3.2.3 micoRNAs as diagnostic biomarkers for CRC

For successful treatment of CRC early detection will provide the best chances for successful treatment. Surgery before metastatic spreading of the disease is considered the only curative form of treatment. The screening methods of colonoscopies and fecal occult blood tests (FOBT) have improved survival rates for CRC by detecting patients and earlier stages of cancer, but there is still much room for improvement and neither test is ideal. While colonoscopies are considered the best screening tool because it can also remove precancerous polyps during the procedure, they are both invasive and expensive, which leads to lower compliance rates. FOBT is less invasive, but also less sensitive and specific. New non-invasive, accurate biomarkers are needed to improve both the accuracy and screening rates for CRC. MicroRNAs are being evaluated for their potential in this area.

3.2.4 microRNA expression as prognostic and predictive biomarkers

Molecular classifiers can serve as prognostic and predictive tools to help stratify cancer patients into appropriate risk groups to aid physicians in making therapeutic decisions. These decisions can include whether or not to provide adjuvant chemotherapy or what types of therapy are appropriate. Expression patterns of microRNAs are associated with both prognosis and therapeutic outcomes in CRC; therefore they have potential as prognostic and predictive biomarkers.

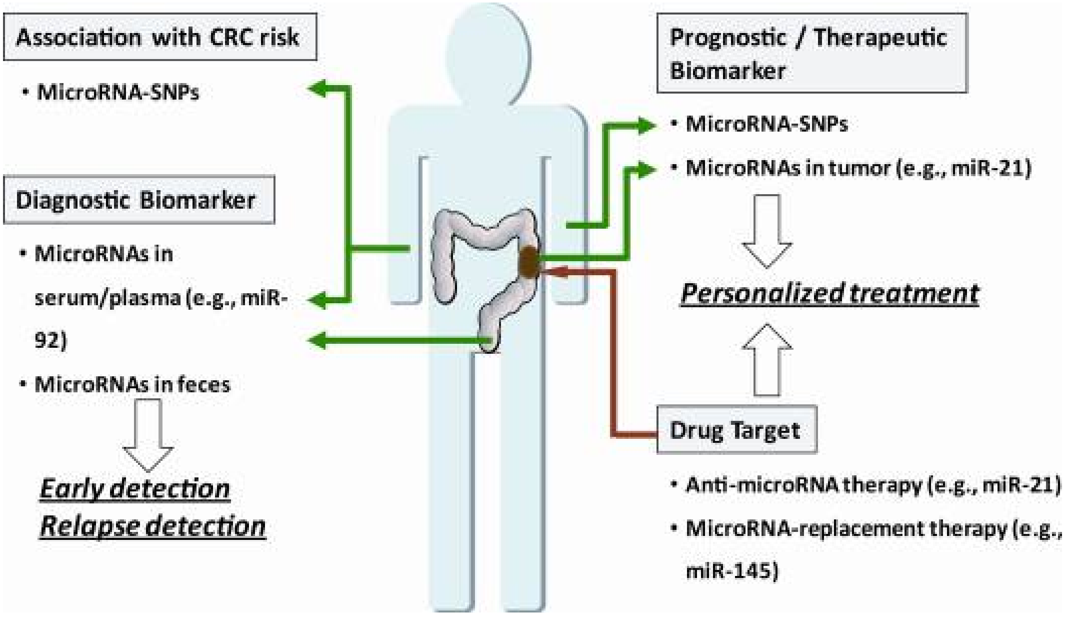

Association with CRC risk
• MicroRNA-SNPs
Diagnostic Biomarker
• MicroRNAs in serum/plasma (e.g., miR-92)
• MicroRNAs in feces
Early detection
Relapse detection
Prognostic / Therapeutic Biomarker
• MicroRNA-SNPs
• MicroRNAs in tumor (e.g., miR-21)
Personalized treatment
Drug Target
• Anti-microRNA therapy (e.g., miR-21)
• MicroRNA-replacement therapy (e.g., miR-145)

Figure 2: Role of miRNA in CRC

3.3 A Machine Learning Approach for the Classification of Kidney Cancer Subtypes Using miRNA Genome Data [13]

In this paper the main objective of the authors were to develop automated tools that can accurately determine kidney cancer subtypes. At the time of publishing the paper, it has been confirmed by researchers from biomedical field that miRNA dysregulation cause cancer. They have built a machine learning approach for the classification of kidney cancer subtypes using miRNA genome data. In this paper through empirical studies it was found that 35 miRNAs possess distinct key features that aid in kidney cancer subtype diagnosis. In this research Neighborhood Component Analysis (NCA) is employed to extract distinct key features from miRNAs and Long Short Term Memory, a type of Recurrent Neural Network, is adopted to classify a given miRNA sample into kidney cancer subtypes.

For this experimental study,the miRNA quantitative read counts data, which was provided by The Cancer Genome Atlas data repository (TCGA). The NCA procedure selected 35 of the most discriminative miRNAs into five subtypes with average accuracy around 95% and Matthews Correlation Coefficient value around 0.92 under 10 runs randomly grouped 5-fold cross-validation, which were very close to the average performance of using all miRNAs for classification.

3.3.1 Data preparation

In this research, kidney cancer RNA-sequence data represented by the miRNA expression that is publicly available on TCGA was used. For kidney cancer three TCGA and two TARGET projects defined the most relevant kidney cancer types as

- High-Risk Wilms Tumor
- Kidney Renal Papillary Cell Carcinoma
- Kidney Chromophobe
- Rhabdoid Tumor
- Clear Cell Sarcoma
- Kidney Renal Clear Cell Carcinoma

3.3.2 Categorization

All kidney cancer cases were considered in which miRNA information was provided. Cases represent the samples taken from patients who had kidney cancer which belonged to one of five different cancer sub-types. For data preparation the following diagram was followed

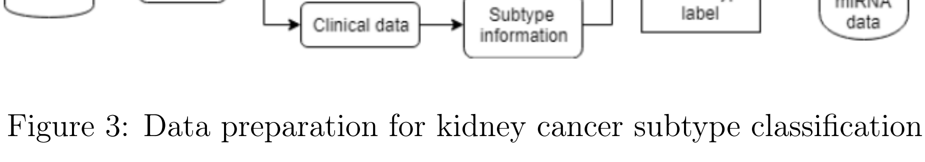


Figure 3: Data preparation for kidney cancer subtype classification

Based on the schematic the following steps were taken.

- Data from TCGA was downloaded and then it was categorized using a MATLAB program.
- First the information related to each case was matched with its miRNA quantification files using the file ID.The information of miRNA read per million for each miRNA was considered in the experiment.
- The miRNA files were then matched with those in clinical data, stored in javascript file, using the case ID. This clinical data provides the record of cancer sub-types and other patient clinical information such as age, sex, and demographics
- The above procedure of preparing the cancer data facilitated automatic classification of kidney cancer sub-types based on the miRNA quantification expression information of the patients.

The most discriminative data features were extracted using NCA algorithm. The authors used all the features to train the model too and found that using all the features generally gave better results. However it was not possible to tell for which features were important for classification.

3.3.3 Model

For classification, LSTM network algorithm with two LSTM layers were used

- The first LSTM layer had 500 neurons
- The second layer had 250 neurons

NVIDIA TITAN X GPU was used for training. Ten runs of five-fold validation was adopted for data analysis which is the procedure followed largely by the Data Analysis Protocal(DAP). The miRNA dataset were randomized for training.

3.3.4 Challenges and Solutions

The challenges that the authors faced were the following

- The dataset was not balanced. To overcome the imbalance in dataset small, random Gaussian noise with zero mean and 0.02 variance were added to the points of those classes with fewer cases, resulting in a balanced dataset [4]. Test samples were excluded from this data augmentation mechanism.

3.3.5 Methodologies applicable in our research

Our research will utilize similar data pre-processing mechanisms. We will also use NCA for feature extraction and LSTM or other variants of LSTM for training our model.

To analyse the result we will use Confusion matrix and Matthews Correlation Coefficient (MCC) [11]. Effectiveness of the selected miRNAs for classification needs to be validated through wet-lab experiments and further clinic studies.

3.4 Deep Learning with Sampling in Colon Cancer Histology [18]

In this paper, the main objective was to apply a deep learning cell identification algorithm to diagnostic images from the colon cancer repository at The Cancer Genome Atlas (TCGA). Here the study shows how experimentally that a cell identification algorithm using deep learning can uncover interesting relationships between tissue morphology and a range of clinical variables and that systematic sampling of tissue regions can improve performance without losing accuracy.

A statistically significant association between morphology and various clinical variables was found in this study; The TNM grading system used in cancer treatment considers tumor penetration, node3s, and metastasis. From this study, we saw Cellularity which is referred to the morphological feature corresponds to the spatial density of the corresponding cell type.

Apart from this cellularity, more features can be calculated using deep learning regarding colon cancer. We may learn about features that include tumor budding, serrated cancers where the colonic glands are of the distinctly serrated form. Jass [8] classified colorectal cancers according to molecular features, Felipe De Sousa et al. [5] reported serrated cancers to have distinct molecular features. So in the future, there is an expectancy that deep learning will help to link the morphological, clinical, and molecular data.

This work already minimized a lot of computational costs. It takes into account a lot of test samples and allows more data to be trained and analyzed if compared with the previous process. Even though there is no guarantee that this approach will always be successful. But further research on finding more features of cell using deep learning will lead to more prominent results.

3.4.1 Methodology and Procedure

In the experiment described in the paper, three Methods were followed which were helpful in the study of identifying cell features which will help in the diagnosis of colon cancer.

1. **Sampling of the set of patches:** Here sampling was employed in the analysis of cases of colon cancer where pathologists were asked to categorize the tissue type at 300 randomly selected points in a dense region of tissue. In histopathology, simple sampling methods cannot be used as it takes spatial dependencies into account. So, two suitable types of sampling were implemented.

   - Random Sampling
   - Systematic Random Sampling

2. **Cell Identification:** Firstly a cell identification algorithm was trained. This algorithm mainly comprises of two convolutional neural networks (CNNs) working in series. The first network detects the cells and the second network simply classifies each cell as epithelial, inflammatory, a fibroblast or as 'others'.

3. **Profile generation:** In the process of Profile Generation images of 1500 cells were hand marked by a pathologist. Cells were classified into four types normal epithelial cells, malignant epithelial cells, inflammatory cells, or fibroblasts. These cell's patches were run through a cell identification algorithm and the accuracies of detection and classification were computed. Both achieved a 65% accuracy on average.

   Both sampling policies are applied using a nominal sample size, two batches were run, and in each batch run the sampling policy were applied to the 142 whole slide images. Different scatter plots and comparisons are analyzed between the two sets of sampling batches. The clinical variable of the data was cross-checked against the four profile features of the cell.

By cross-checking, all the data different combinations of the profile features were found resulting in unique problems like Mucinous carcinomas were associated with fewer inflammatory cells than were non-mucinous carcinomas, and for metastasis, residual tumor, and venous invasion were related to lower numbers of epithelial cells.

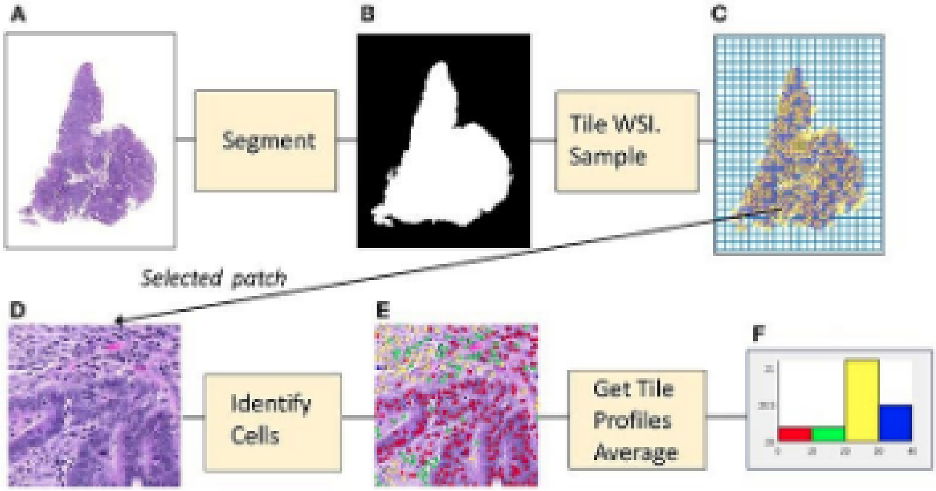


Figure 4: In the figure above the process is summarized on how the cell feature is analyzed

From this paper we can see how images can be analysed to identify CRC. Our research will explore miRNA expression for CRC so we might not used the methodologies introduced here. We could use random sampling and systematic random sampling for data augmentation in our procedure.

3.5 Identification of 12 cancer types through genome deep learning

In this paper [19] the authors create a new neural network called Genome Deep Learning. Deep neural network, a high level abstraction algorithm very competent in finding patterns in large sets of data, was used as a classifier to diagnose patients with 12 different types of cancer. Data that was used to train and test the classifier were point mutation data, which denote the genomic variation between a harmful allele of the principal gene and their healthy variants. Mutations in oncogenes play more of a role in cancer development compared to any other genes and therefore those were chosen as a principal point of focus, some examples would be BRCA1 and BRCA2 for the development of breast cancer. The process outlined in the paper includes data collection and preprocessing, model training (deep neural network in TensorFlow environment) and evaluation.

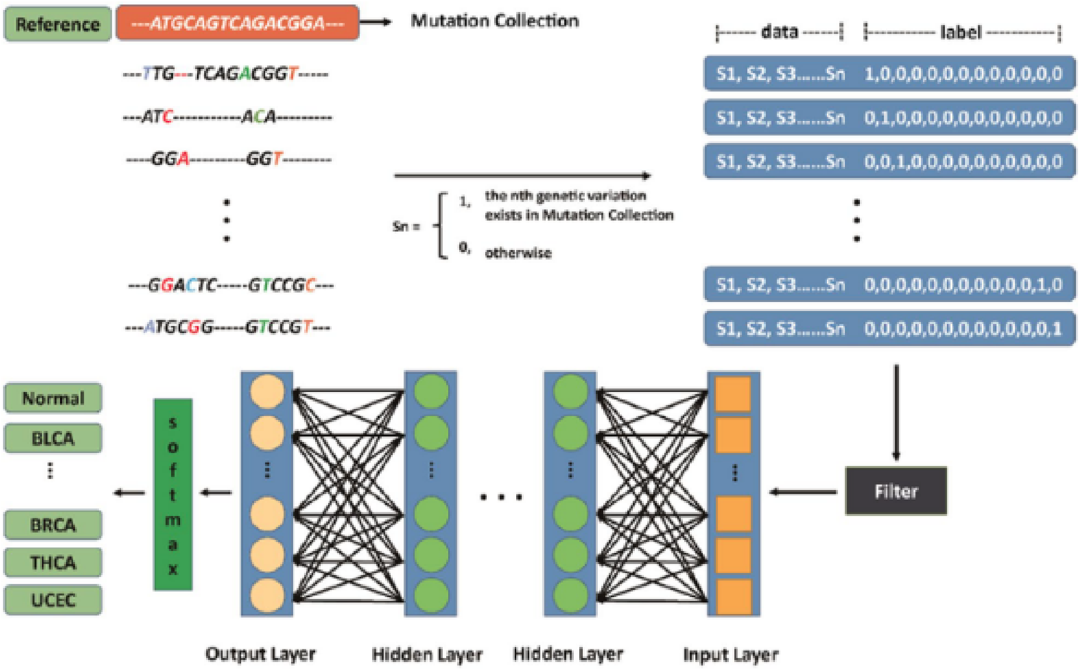

Reference
---ATGCAGTCAGACGGA---
Mutation Collection
---TTG---TCAGACGGT-----
---ATC----------ACA---------
----GGA---------GGT--------
---GGACTC-----GTCCGC---
---ATGCGG-----GTCCGT---
Sn = 1, the nth genetic variation exists in Mutation Collection
0, otherwise
data
label
S1, S2, S3......Sn 1,0,0,0,0,0,0,0,0,0,0,0,0
S1, S2, S3......Sn 0,1,0,0,0,0,0,0,0,0,0,0,0
S1, S2, S3......Sn 0,0,1,0,0,0,0,0,0,0,0,0,0
S1, S2, S3......Sn 0,0,0,0,0,0,0,0,0,0,0,1,0
S1, S2, S3......Sn 0,0,0,0,0,0,0,0,0,0,0,0,1
Filter
Normal
BLCA
BRCA
THCA
UCEC
softmax
Output Layer
Hidden Layer
Hidden Layer
Input Layer

Figure 5: Architecture of GDL (Genomic Deep Learning)

The genomic data of patients were compared to healthy alleles to generate point mutation files and the most significant point mutations were chosen as the dimensions of the input data. Through this model, the authors were able to achieve the accuracy, specificity and sensitivity values of 94.70%, 97.30% and 85.54%, respectively. The method outlined in the paper is able to diagnose a patient irrespective of the stage the cancer is in and therefore fare better in diagnosis compared to traditional means when the cancer is in a nascent stage. A limitation the proposed methodology faces is that it does not take into account other factors such as age, sex, trascriptome and proteome data, which may contribute to accurate diagnosis. Another limitation is the number of cancers that are possible to diagnose is still very low and there is room for further research and development.

3.6 Deep Residual Learning for Image Recognition

Deep neural networks are actually complicated to train. So the residual learning is used here for image classification. The residual Networks are easier to optimize and they can get accuracy to a higher level. Using this network of the ImageNet dataset we were able to get results for a depth of 152 layers (8 times deeper than VGG) with a very lower complexity with an error of only 3.2%.
The deep convolutional neural network has breakthroughs for Image Classification. However, it faced some complications like the vanishing gradient problem which hampers the convergence. When deeper networks are able to start converging a degradation problem has been exposed, with increase in depth the accuracy falls and degrades rapidly. This shows not all systems are similar to optimize, and thus the deep residual learning framework is introduced for the degradation process.

Residual Network:
The residual learning is for every few stacked layers.

$$y = f(x, \{W_i\}) + W_s x$$

It is the equation for the residual mapping. x and y are the input and output vectors. The RELU and the biases are omitted for simplifying notation. The operation '*f+x*' was performed by shortcut connection and element-wise addition. No extra parameter or computation complexity increases for the shortcut connection. The form of the residual connection '*f*' is flexible which has multiple layers, but if it only has one layer then it is similar to a linear layer.The function

$$f(x, W_i)$$

can represent multiple convolutional layers. The element-wise addition is performed on two feature maps, channel by channel.
For experiments two networks are used:

- *Plain Network*
  It has convolutional layers mostly have 3x3 filters and the network has global average pooling layers and 1000 way fully connected layers with softmax.

- *Residual Network*
  Based on the plain network, shortcut connections are applied which turns the network into its counterpart residual version.

Experiments on both ImageNet Classification and also with the CIFAR-10 data set are done for the two different networks .

For ImageNet Classification ,with the plain network it is seen that when a deeper 34- layer network is applied rather than the 18-layer, the validation error increases and it has higher training error as well. However, for Residual networks the 34-layer network gives lower error than the 18-layer network and the accuracy keeps increasing with increasing the number of layers. The 152 layer gives even better results than the 34-layer network.

For CIFAR-10, using a plain network it is seen that the error keeps increasing with more number of layers, and at the same time lots of parameters are required. However, using the residual network the error decreases for higher layer networks and even less parameters are required compared to the plain network. Though there is a limitation, when more than 1000 layers are used, the optimization becomes difficult and the accuracy starts to fall due to overfitting. So for such a data set it is unnecessary to use more than 1000 layers.

Using the residual network it is also seen that it has good generalization performance on other recognition tasks such as PASCAL or in MS COCO. Faster R-CNN detection methods can be obtained. Research on this particular field is still going on where this residual network can be used to improve the object detection techniques even further.

4 Proposed Methodology

A typical machine learning algorithm starts with feature selection, though deep learning algorithms can also be designed to handle raw data [21]. With regard to feature selection, it was demonstrated in [22] that NCA is an effective method for selecting significant feature points for high-dimensional data. This method is a nearest neighbor-based feature weighting algorithm. As a feature selection tool, the NCA method was successfully tested on several microarray datasets for various cancers, such as colon cancer, brain tumor, leukemia, lung cancer, and prostate cancer [22].In this research, the NCA algorithm was adopted for selecting high-rank features form miRNA data. We will also consider using artificial neural network to train our model

4.1 Neighborhood Component Analysis

Let us consider a multi-class classification problem. Let c be the number of classes and n be the number of observations. Then a given training set can be described as follows [22]:

$$S = \{(x_i, l_i), i = 1, 2, 3, .., n\} \tag{1}$$

where $x_i \in R^i$ are the featire vectors and $I_i \in \{1, 2, ..., c\}$ are the class labels. Let $f : R^p \to \{1, 2, ..., c\}$ be the classifier to be trained.

Consider a randomized classifier that picks a reference point randomly, *Ref(x)* then labels x using the label of the randomly selected reference point *Ref(x)*. Choice of reference point is based on some probability, which is called the selection probability. The probability $P($*Ref(x)* $= x_j|S)$ will be higher if the reference point of x, x_j, is closer to x, as measured by the distance function

$$d_w(x_i, X_j) = \sum_{r=1}^{p} w_r^2 |x_i r - x_j r| \tag{2}$$

Where w_r for r= 1,2..,p are feature weights. Assuming that the selection probability is direct proportional to k($d_w(x_i, x_j)$), where k is a kernel or similarity function, such that it produces large values when $d_w(x_i, x_j)$ is small.Since the reference point is chosen from the set, the sum of $P($*Ref(x)* $= x_j|S = 1)$ for all j [22]. Thus we consider the following probability P

$$P(Ref(x) = x_j|S) = \frac{k(d_w(x_i, x_j))}{\sum_{j=1}^{n} k(d_w(x_i, x_j))} \tag{3}$$

This is a classifier using the strategy of leave-one-out (Training on all points excluding a single point). The probability that point x_j is picked as the reference point for x_i is

$$P_{ij} = \frac{k(d_w(x_i, x_j))}{\sum_{j=1, j\neq i}^{n} k(d_w(x_i, x_j))} \tag{4}$$

$$P_i = \sum_{j=1, j\neq i}^{n} p_{ij} l_{ij} \{ \tag{5}$$

where $l_{i_j} = 1$ if the $y_i = y_j$ or 0 if the condition is not satisfied. Where p_i is the average leave-one-out probability of correct classification of the observation i using S^i. We can express the probability of correct classification by the randomized classifier as

$$F(w) = \sum_{i=1}^{n} p_i - \lambda \sum_{r=1}^{p} w_r^2 \tag{6}$$

where λ is the regularization parameter, and F(w) depends on the weight vector w. The Neighborhood Component Analysis procedure tries to find the maximum F(w) with respect to w. Many of the weights in w will vanish by regularization. We can find the vector w by minimizing (6) given lambda.

4.2 LSTM

LSTM is one type of Recurrent Neural Network that deals mostly with sequential data. In LSTM to retain memory cell state is used. The cell state is similar to production chain, the parameter flows straight forward, but some linear processes, such as addition and multiplication, will interact. The state of the cell depends on the interactions, and if there are no interactions, it will flow along without changes. LSTM will add or remove information to the cell state through gates which are structures that allow optional information to cross. Gates are implemented using sigmoid functions which produces two decisions either 0 or 1 assuming that 0 will block information flow and 1 will allow it. Three of these gates are available in LSTM, which determines the final cell state.

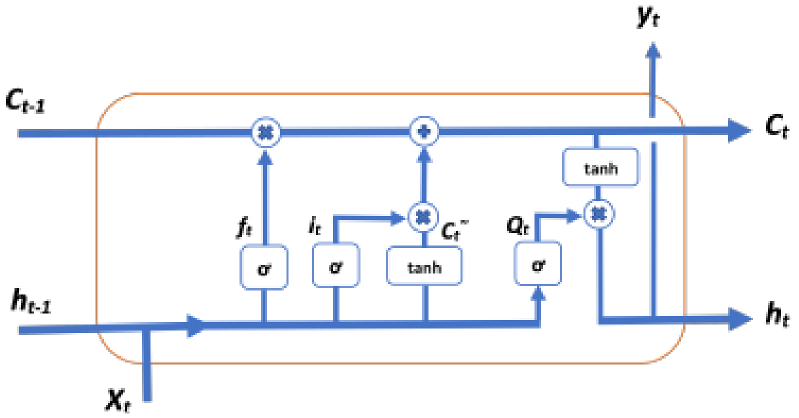


Figure 6: One Long Short-Term Memory block

The block or neuron shown here in Figure 3 is described by the following functions

$$f_t = \sigma(W_f.[h_{t-1}, x_t] + b_f) \tag{7}$$

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i) \tag{8}$$

$$\widetilde{C}_t = \tanh W_C[h_{t-1}, x_t] + b_C \tag{9}$$

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t \tag{10}$$

$$Q_t = \sigma(W_Q.[h_{t-1}, x_t] + b_Q) \tag{11}$$

$$h_t = Q_t * \tanh C_t \tag{12}$$

where

- f_t is the activation vector of the forget gate
- σ is the sigmoid function
- W is weight matrices to be learned during training,
- x_t is the input vector to the LSTM unit
- b is the bias vector parameters to be learned during training,
- i_t is activation vector of the input gate,
- C_t is cell state vector,
- Q_t is activation vector of the output gate and
- h_t is output vector of the LSTM unit.

4.3 Artificial Neural Network

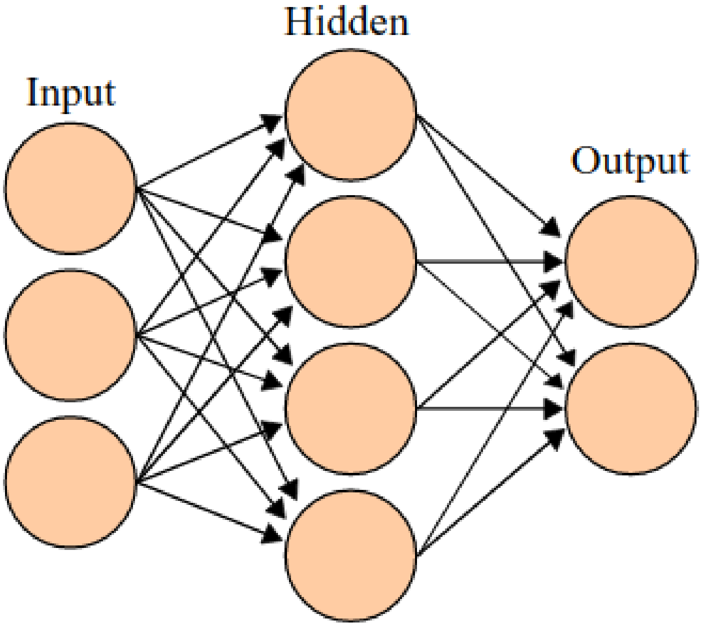


Figure 7: Neural Netowrk Architecture

Artificial Neural Networks are layers of neurons connected together in sequence. The idea behind these types of architecture is to capture non linear patterns in the data. It usually contains three types of layers input layer, hidden layer and output layer. The input layers are used to take the input of the data. The hidden layer are used to introduce non linear activation functions. This functions transform the data non linearly. So that they can be later separated linearly. The output layer finally gives out the probability to which the classes belong to.

4.4 Ensembles Neural Network

Methods:
Methods employ the potential of many machine learning models and then use a voting function to evaluate labels of data. The primary advantage of ensemble learning is that it tries to solve the high variance problem endemic to deep learning models in general when dealing with complex models. The series of learning models or classifiers that are used in the primary phase are called base learners. Our goal in the primary phase is to improve the diversity of output, for this a variety of different types of learners are used for classification or the same type of learner may be used but with different subsets of the data set being used to train the models. When we receive the output from each of these base learners a voting or averaging technique is used to obtain the final label. Two different methods of ensemble learning are the most prominent.

Bootstrap Aggregating:
Also known as bagging, in this method from the training data set, random samples with replacement are chosen and fed to the individual base learners. After training the models are run on a test data set. The results or output classes from these individual models are aggregated for which there exist various methods as discussed previously. In bagging, all the individual base learners can be trained simultaneously.

Boosting:
In this process each tuple or instance in the training dataset has a weight associated with it. These weights represent the probability of being chosen when random selection occurs for training the individual base learners. For each training unit we randomly select from the dataset and then train the model. After training the entire dataset is used as a test set and the labels are determined accordingly. For those instances of the dataset the labels were incorrect, their weights are increased, this has the implication that when the next base learner is to be trained, the instances with the higher weights will have a higher probability of being chosen in random selection.

4.5 ResNet

As our model gets deeper and deeper we are proposing a neural network with residual block for skip connections. Skip connections enable models to add the input the activation output in order to retain lost information. This has been used commonly in image classification and it is a common base network in object detection.

In figure 8 the skip connections are denoted by curved arrows. It can be observed that a part of input is passed through the convolutional layers and the other part is skipped and added with the output of the convolutional layers.

4.6 Data Preparation

In this step we are going to download all the publicly available miRNA expression data in TCGA from the following projects

- Acute Myeloid Leukemia,Bone Marrow
- Infiltrating duct Carcinoma, Breast
- Clear cell adenocarcinoma, Kidney
- Serous cystadenocarcinoma, Ovary

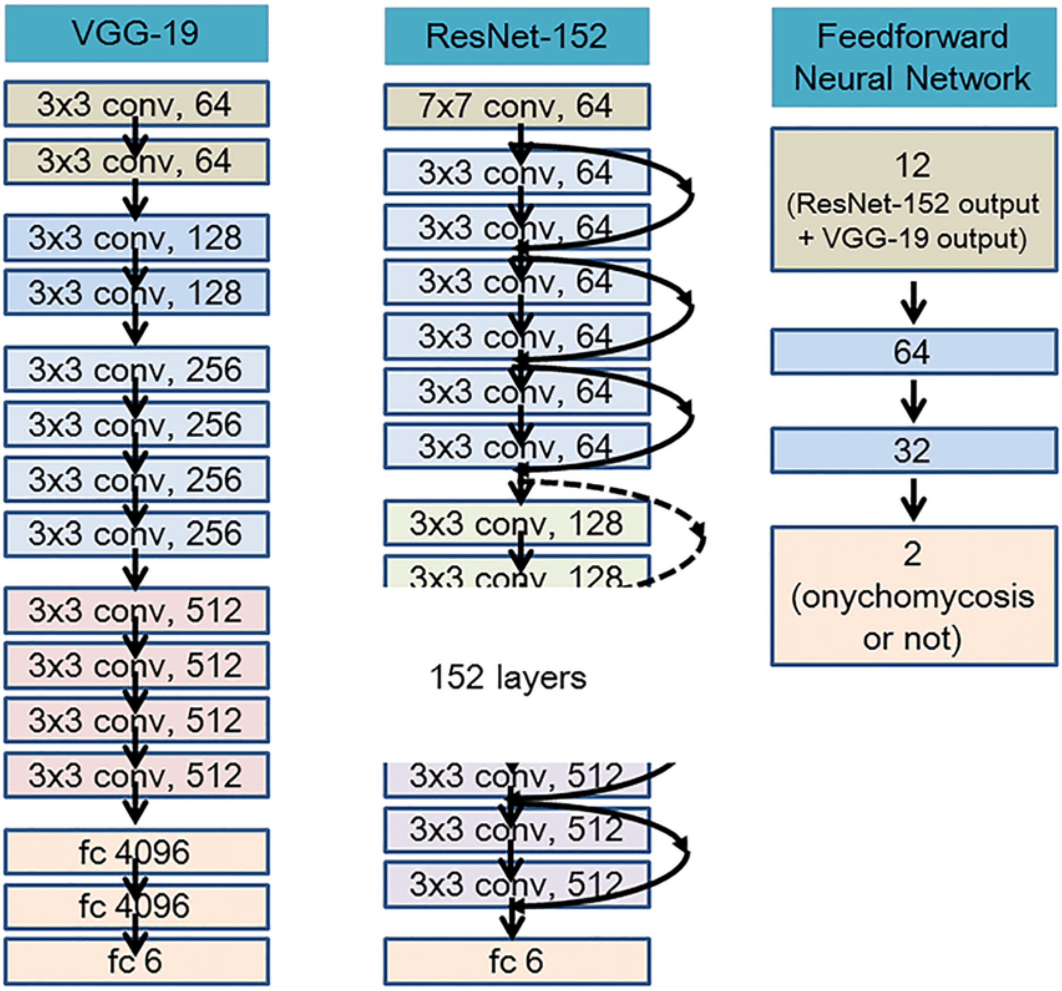


Figure 8: Comparison of ResNet with different architectures

- Malignant melanoma, Skin
- Squamous Cell Carcinoma, Lung
- Papillary Adenocarcinoma, Thyroid Gland
- Papillary Adenocarcinoma, Kidney

After downloading the data we will follow the following schematic diagram to preprocess our data.

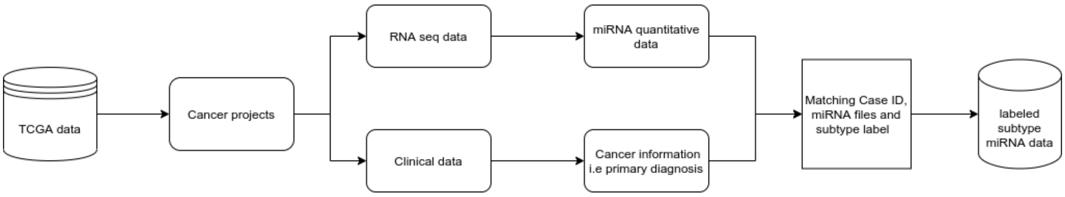

TCGA data
Cancer projects
RNA seq data
Clinical data
miRNA quantitative data
Cancer information i.e primary diagnosis
Matching Case ID, miRNA files and subtype label
labeled subtype miRNA data

Figure 9: Data preprocessing architecture

For our initial training phase we have downloaded data from the following 5 tissues

- Bone Marrow
- Breast
- Kidney
- Prostate Gland
- Ovary
- Lung
- Thyroid Gland
- Skin

These data consists of miRNA quantification file for all the cancer subtypes of the 5 tissues. The following diagram shows the distribution of the miRNA quantification files data

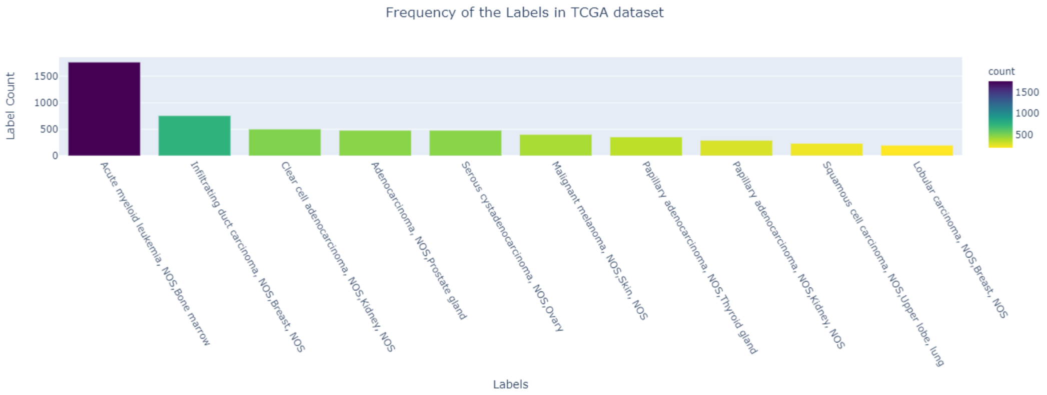


Figure 10: Data distribution of all samples

Cancer Types	Number of Samples
Acute Myeloid Leukemia,Bone Marrow	301
Infiltrating duct Carcinoma, Breast	301
Clear cell adenocarcinoma, Kidney	301
Serous cystadenocarcinoma, Ovary	301
Malignant melanoma, Skin	301
Squamous Cell Carcinoma, Lung	235
Papillary Adenocarcinoma, Thyroid Gland	301
Papillary Adenocarcinoma, Kidney	291

Table 1: Sample Counts

Train Samples	1749
Test Samples	583

Table 2: Train and test split for the experiment

Then we selected only 300 samples from the top 8 classes given in the top diagram. Which gives an equal distribution of data
Then we fed this data to our models

4.7 Model

For our deep learning model first we are going to extract the most distinguishable features using NCA as described above. Then we will pass it through our LSTM network for training and evaluate the performance of our model using confusion matrix and Matthew's coefficient [11]. We are also going to use ANN model for training this model. We have created a ResNet variant where instead of using Convolutional layers we are using dense layers with residual blocks. We are also creating an ensemble method based on 5 models trained on different architectures

This diagram gives an overview of our procedure.

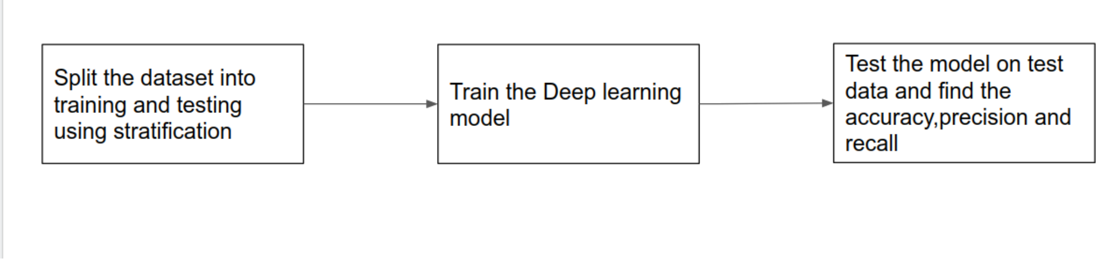

Split the dataset into training and testing using stratification
Train the Deep learning model
Test the model on test data and find the accuracy,precision and recall

Figure 11: Model training architecture

In order to train the deep learning model it is imperative that we split the data in balanced manner across all the classes so that there is no imbalance on our dataset that we are going to feed to our model. We must also keep a portion of data separate which won't be seen by the model in the training phase. This data will be used to test our model on unseen data, so that we can apply it in real life scenarios.For this reason, we have split the dataset into 75% and 25% for train dataset and test dataset respectively. From the following diagram it can be observed that each classes have the same number of sample counts. Before passing the data to the model we normalize the data to reduce variance and make mean 0.

4.7.1 ANN

After splitting our data, we have trained our ANN model. Neural Networks are inspired from biological neurons. At each layer there are nonlinear activation functions. These functions introduce non linearity to capture the non linear patterns in the data.Our model used 5 dense layers and between each layer we used batch normalization and dropout to introduce regularization to ensure that the model doesn't over fit the data. Otherwise the model will fail to generalize on unseen data.

The diagram in figure 12 shows our Neural Network architecture

4.7.2 LSTM

Our LSTM model has 3 LSTM units and 2 Dense layers. It uses an embedding layer to embed the features so that it can be captured by the model. The idea is to capture sequential information such as context. In each time step the the information from previous time step is either retained or forgotten. Then, it takes the input from the current time step. Finally it calculates the and updates the state of current cell and gives the output

The diagram in figure 13 shows our LSTM architecture

4.7.3 CResNet

CResNet is a variant of the popular deep learning architecture called ResNet. The difference between these two networks is that instead of using convolutional layers, dense layers are used. It consists of Residual Block which performs the skip connection. The diagram in figure 14 represents our CResNet architecture

Inside ResBlock there are two Dense layer and two batch normalization layer. The batch normalization layer normalizes the activation output in each dense layer.Normalizing the inputs to the layer has an effect on the training of the model, dramatically reducing the number of epochs required. It can also have a regularizing effect, reducing generalization error much like the use of activation regularization. The activation layer applies non linear activation function on the output of the final Batch Normalization layer to get the non linear effect. Finally the input of Resblock is added with the activation output.

Finally our CResNet model consists of 4 dense layers and 2 ResBlock layers. Each layer has a Relu activation function. The final layer has sigmoid function which gives us the probability for which each classes belong to.

4.7.4 Ensemble

Deep learning models compared to traditional machine learning models can offer an increased amount of flexibility and can be scaled in proportion based on the amount of available training data. However since deep learning models have large number of layers they tend to over fit which might cause high variance. Generally that means, the model performs well on training data but fails to generalize well on unseen data.

So the alternative way to tackle is to introduce ensemble learning. In this approach instead of training a single model multiple models are trained. The results of multiple mobiles are combined

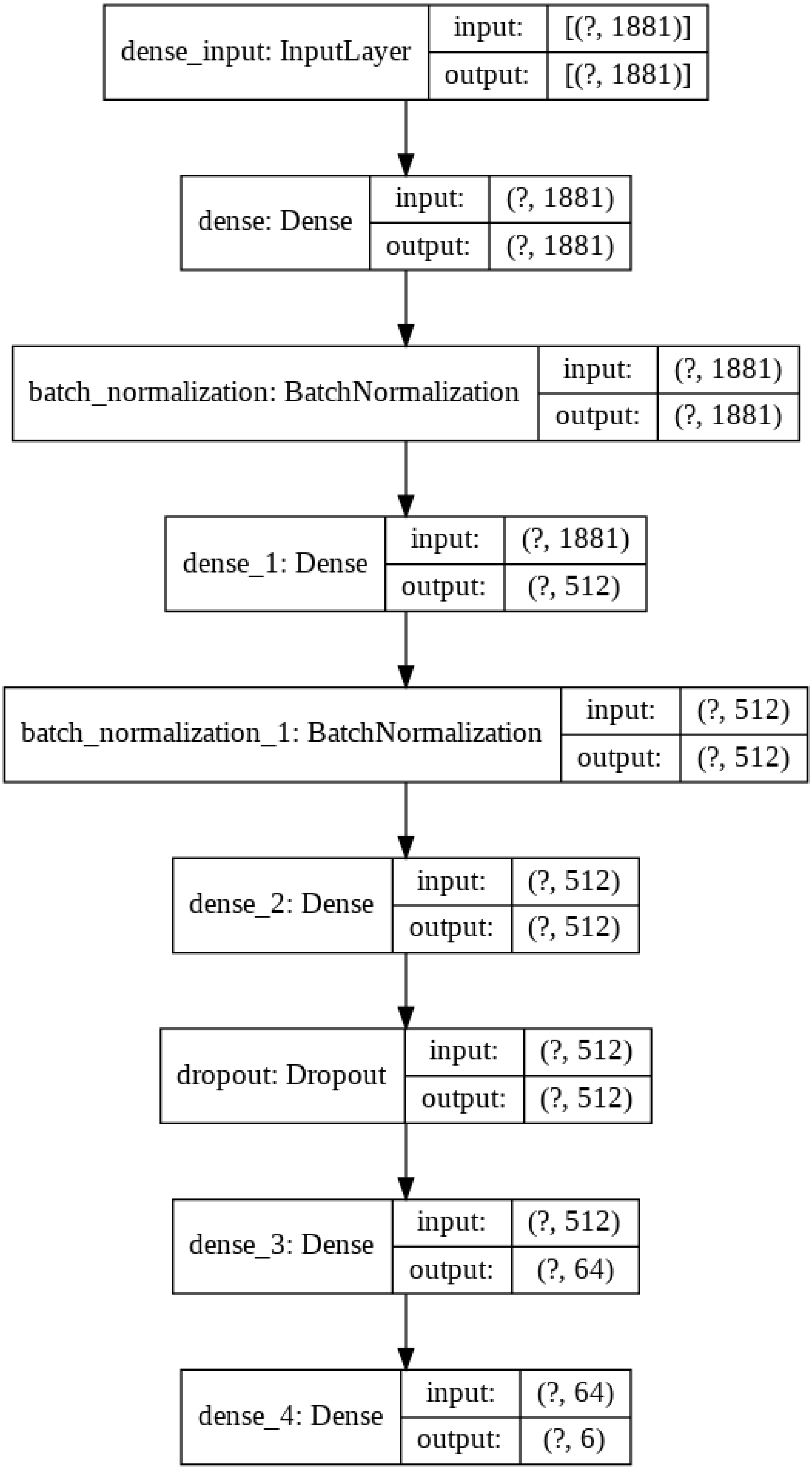

dense_input: InputLayer | input: [(?, 1881)] | output: [(?, 1881)]
dense: Dense | input: (?, 1881) | output: (?, 1881)
batch_normalization: BatchNormalization | input: (?, 1881) | output: (?, 1881)
dense_1: Dense | input: (?, 1881) | output: (?, 512)
batch_normalization_1: BatchNormalization | input: (?, 512) | output: (?, 512)
dense_2: Dense | input: (?, 512) | output: (?, 512)
dropout: Dropout | input: (?, 512) | output: (?, 512)
dense_3: Dense | input: (?, 512) | output: (?, 64)
dense_4: Dense | input: (?, 64) | output: (?, 6)

Figure 12: Neural Network

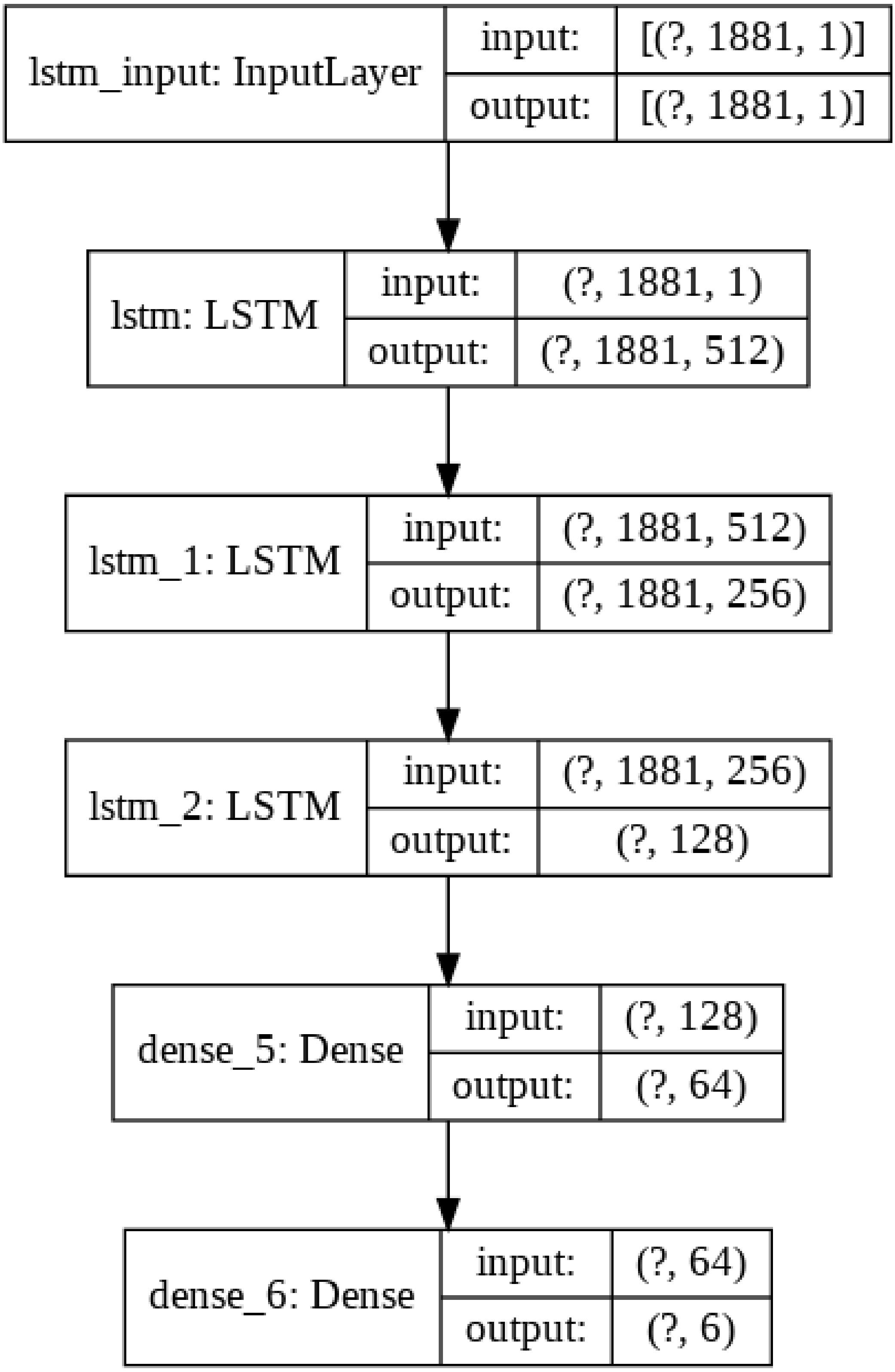


Figure 13: Lstm

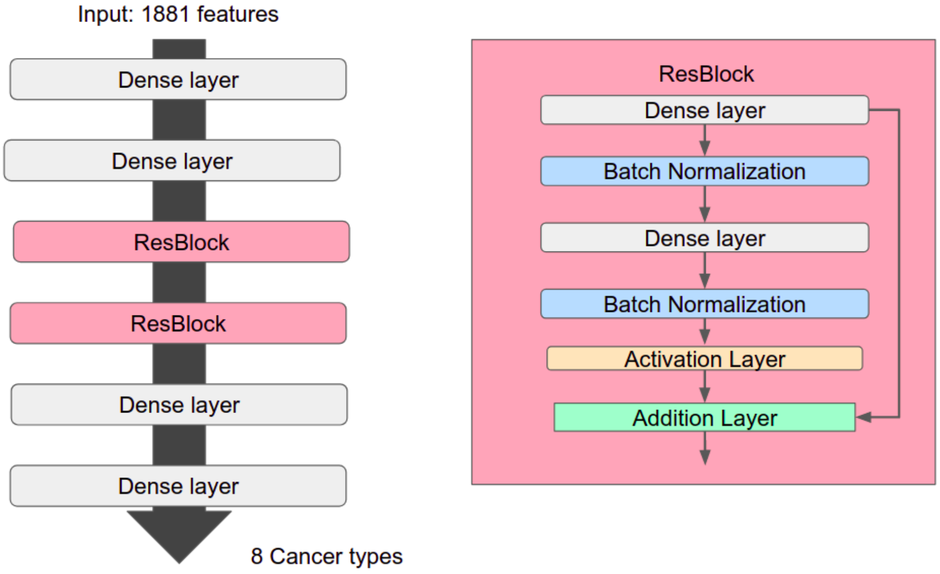


Figure 14: CResNet

to give predictions. This not only reduces overfitting but it can also result in predictions that are better than any single model. The diagram in figure 16 shows our model Architecture

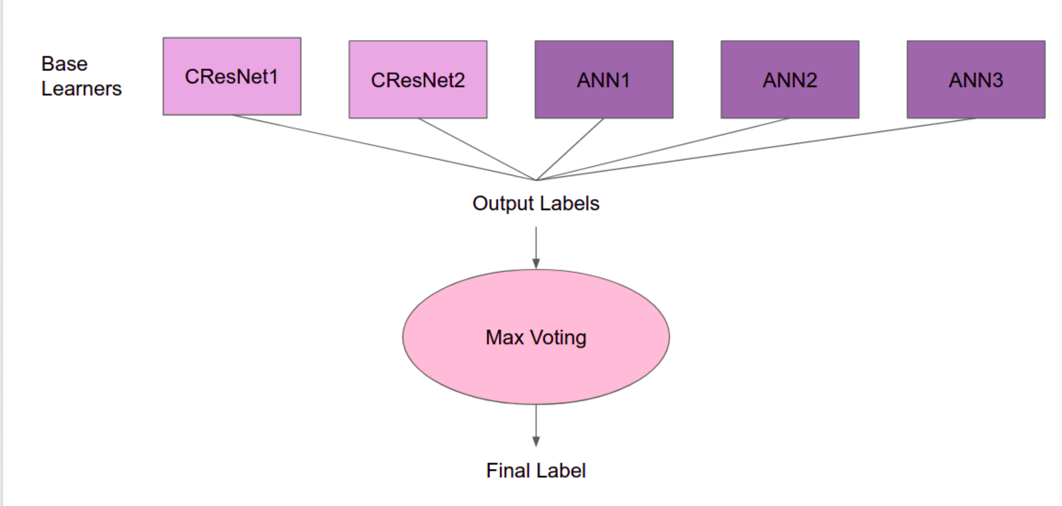


Figure 15: Ensemble method using model averaging

4.8 Activation and Optimizer

For the activation functions two activation function were used. Relu was used to introduce non linearity. Softmax activation was used to calculate the probability of the class based on the output of the final layer.For optimization Adam was used

4.8.1 Relu

Relu is a non-linear activation function that is used to introduce non linearity in the deep learning network.It takes the maximum of the given value and 0.

$$Relu(0, x) = max(0, x) \tag{13}$$

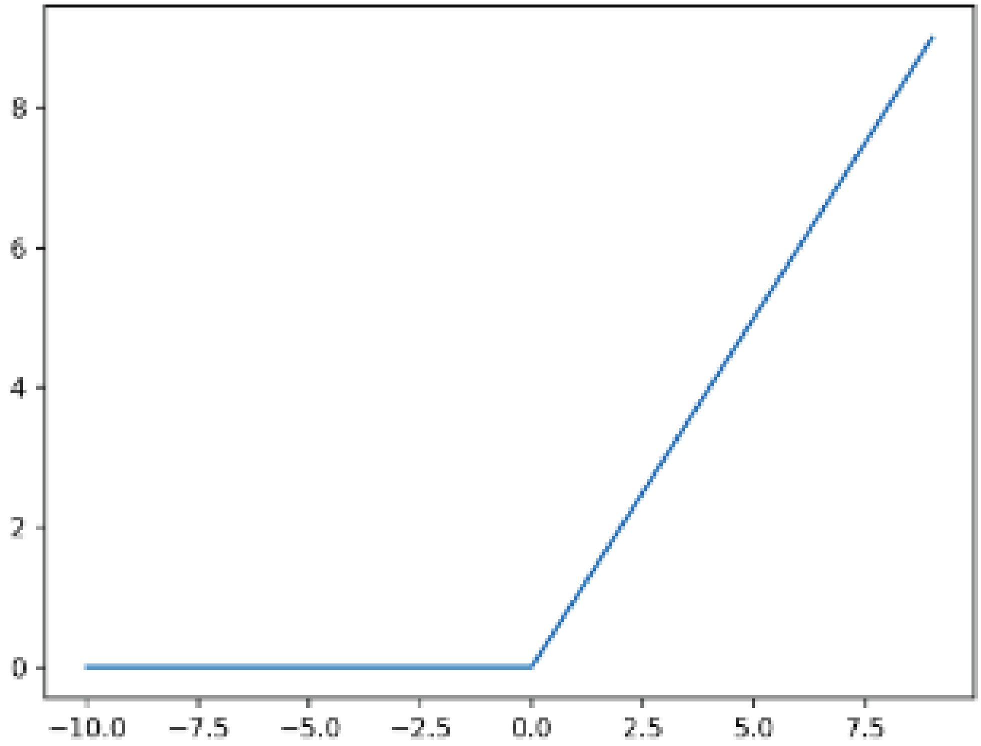


Figure 16: Relu

4.8.2 Softmax

It is used in multi classification problems. It takes the logits(output of the last layer) and pass it through a softmax function which classifies to which class these features represent

$$Softmax = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_i}} \tag{14}$$

4.8.3 Adam

Adam optimization algorithm is an extension to stochastic gradient descent that has recently seen broader adoption in the deep learning world. Adam stands for adaptive moment estimation. On non-convex optimization problems Adam offers the following benefits

- Straightforward to implement
- Computationally efficient compared to other optimization algorithms
- Little memory requirements
- Well suited for problems that are large in terms of data and/or parameters

4.9 Matthew's Correlation Coefficient(MCC)

Matthew's Correlation is a machine learning metric which is used to measure the quality of classification. It is a widely used metric in the field of Biometrics.It always yield value in between -1 to 1. It basically measures the correlation between the true and predicted values. The confusion matrix is the easiest way to represent our model results as it uses the coefficient which takes into account the true and false positive and negative values.This gives a balance in our measures even though the classes are of different size The equation is shown below

$$\frac{TP.TN\text{-}FP.FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{15}$$

4.10 Training

For training our model we have used TensorFlow's Keras API. We have used the following configuration for our model ANN, CResNet, Ensemble and LSTM. We have trained all the models on google colab.

- Optimizer: Adam
- Learning rate: 0.001
- Epochs: 200
- Batch size: 64
- Framework : TensorFlow
- GPU : Tesla K80
- RAM: 13 GB

Since we are going to use Stochastic gradient descent like Adam optimization our loss will fluctuate a lot. Because a single batch cannot capture the entire variance of the dataset. Which means a particular batch might increase the loss and the other batch might decrease the loss.

The diagrams in figure 17 ,figure 18 and figure 19 shows the accuracy and loss of both the ANN model and LSTM model

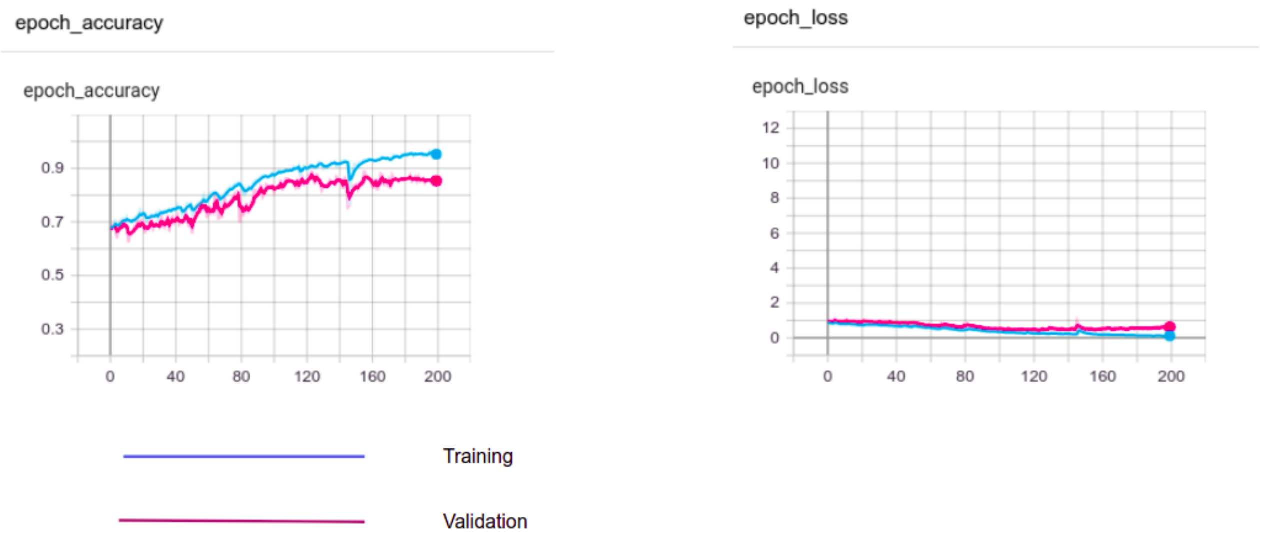


Figure 17: Accuracy and Loss of LSTM

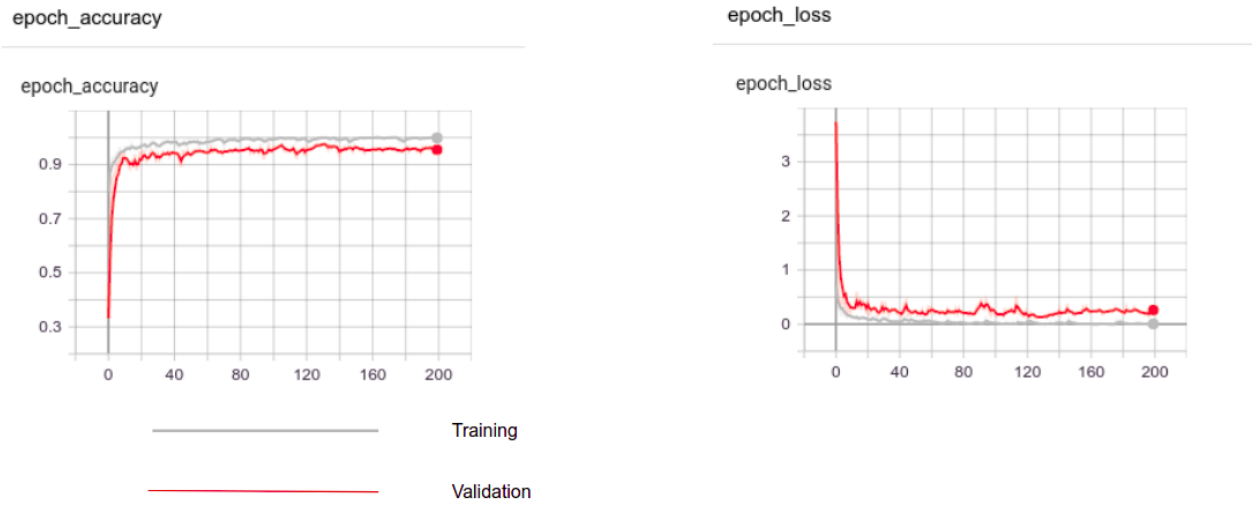


Figure 18: Accuracy and Loss of CResNet

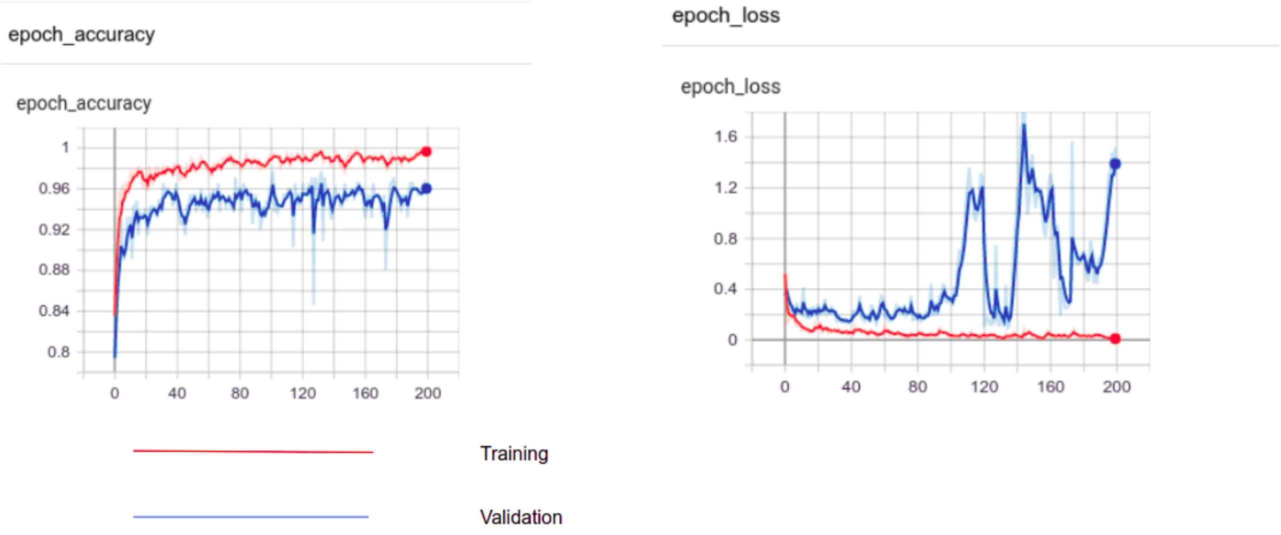


Figure 19: Accuracy and Loss of ANN

5 Result Analysis

For our result analysis we will be analyzing the performance of our models in 3 ways

- Metrics Comparison
- Confusion Matrices
- Tables for calculating Mathew's correlation coefficient

5.1 Metrics Comparison

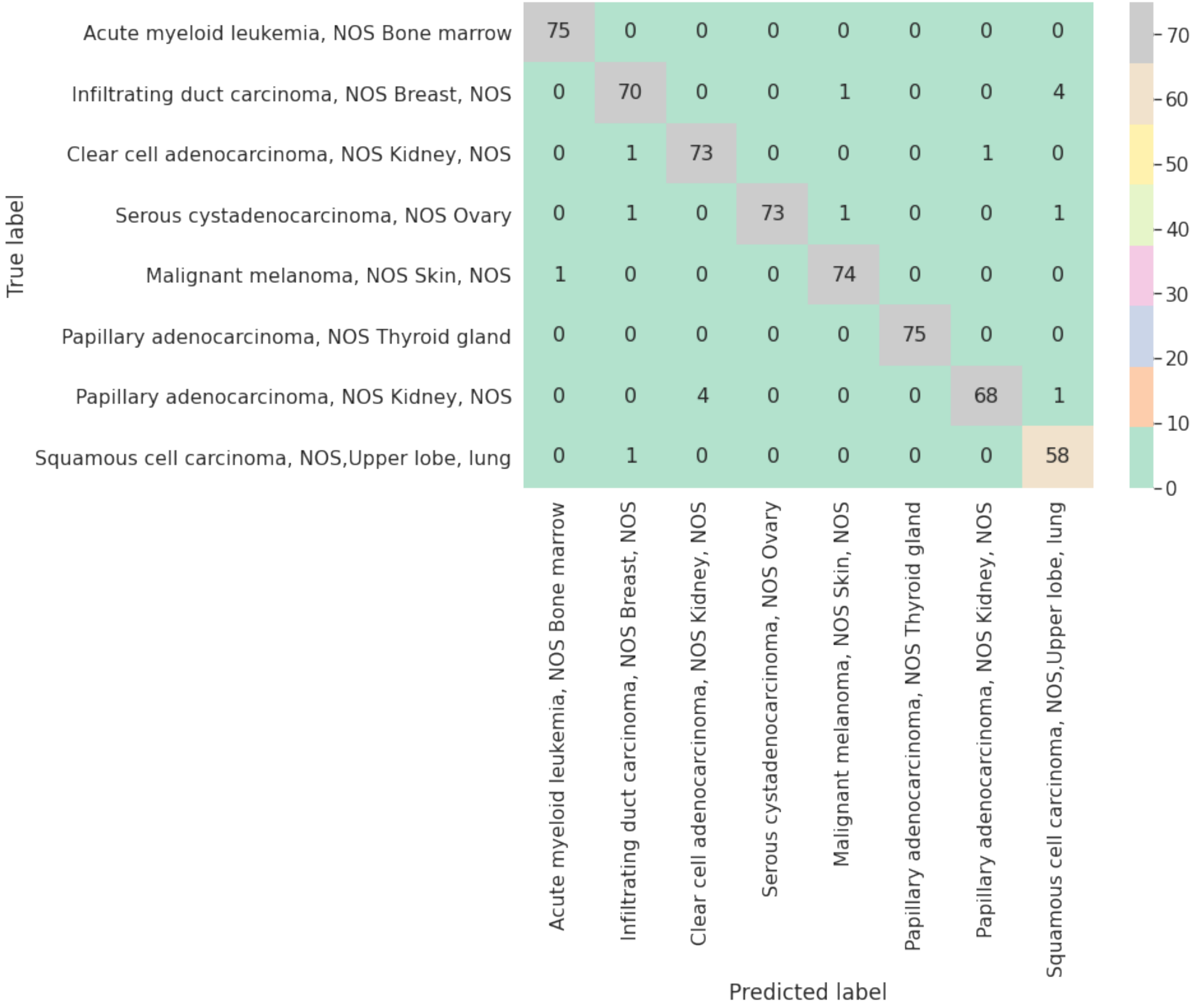
In Metrics Comparison we have combined the accuracy, loss and MCC value of the 4 types of model we trained to demonstrate which model has the best performance. From the table it can be seen that the architecture we have proposed using CResNet and ANN, outperforms the LSTM architecture proposed by the paper of Kidney Cancer Classification [13]

	Loss	Accuracy	MCC
Artificial Neural Network	0.1864	97.07%	0.9625
LSTM	0.4079	87.48%	0.8596
CResNet	0.2876	97.77%	0.9745
Ensemble	N/A	95.03%	0.9439

5.2 Confusion Matrices

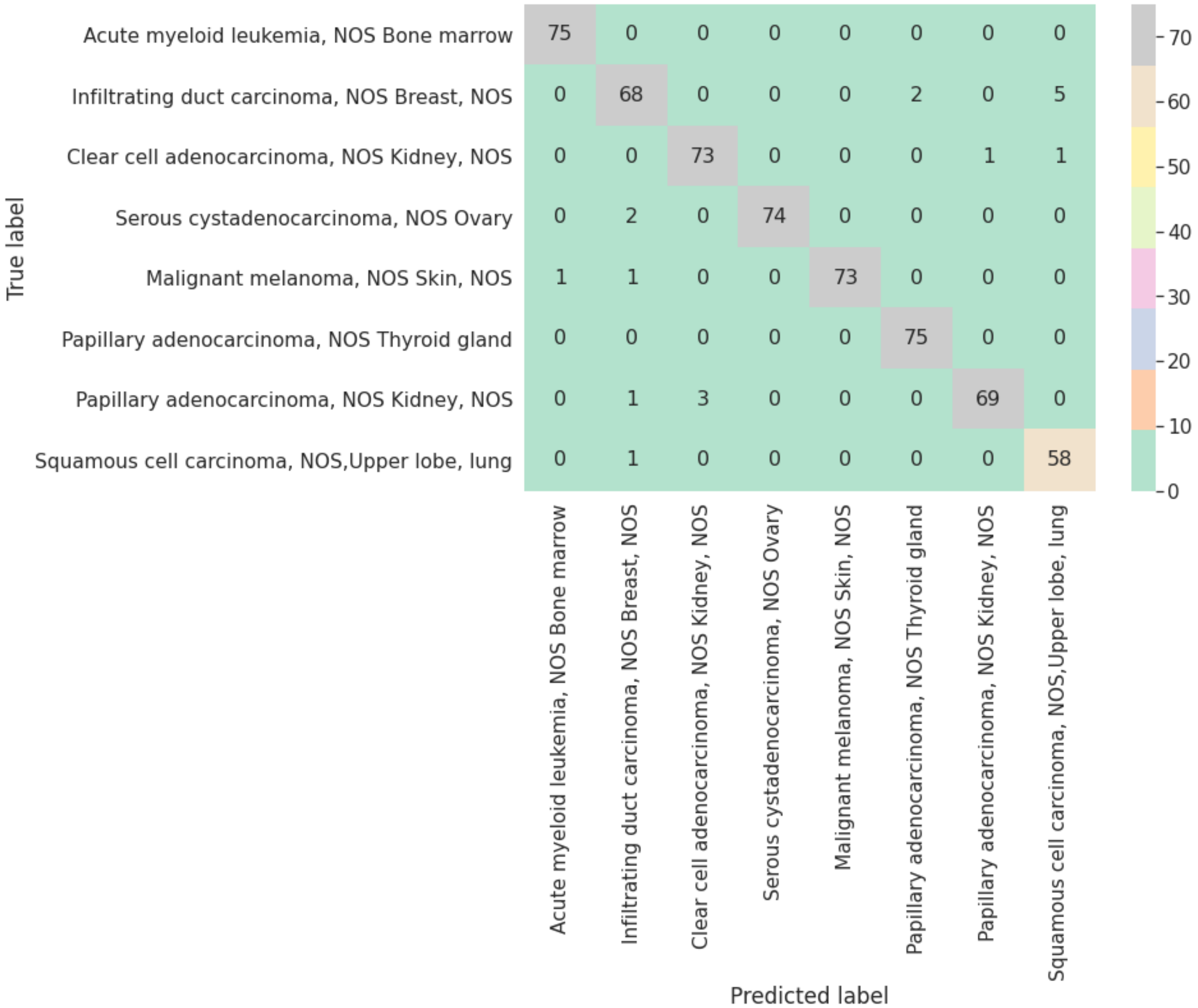
One of the most popular way to demonstrate the classification of a machine learning model or deep learning model is with the use of Confusion matrix. A confusion is a matrix which consists of the actual labels denoted by the rows and predicted labels denoted by the column. From this matrix it is possible to calculate true positive, false positive, true negative and false negative. These metrics can be further used to calculate accuracy, MCC, recall, precision, F1-score etc.

The following figures 20, 21 ,22, 23 represents our confusion matrices of the models we have trained for ANN, CResNet, Ensemble and LSTM models respectively



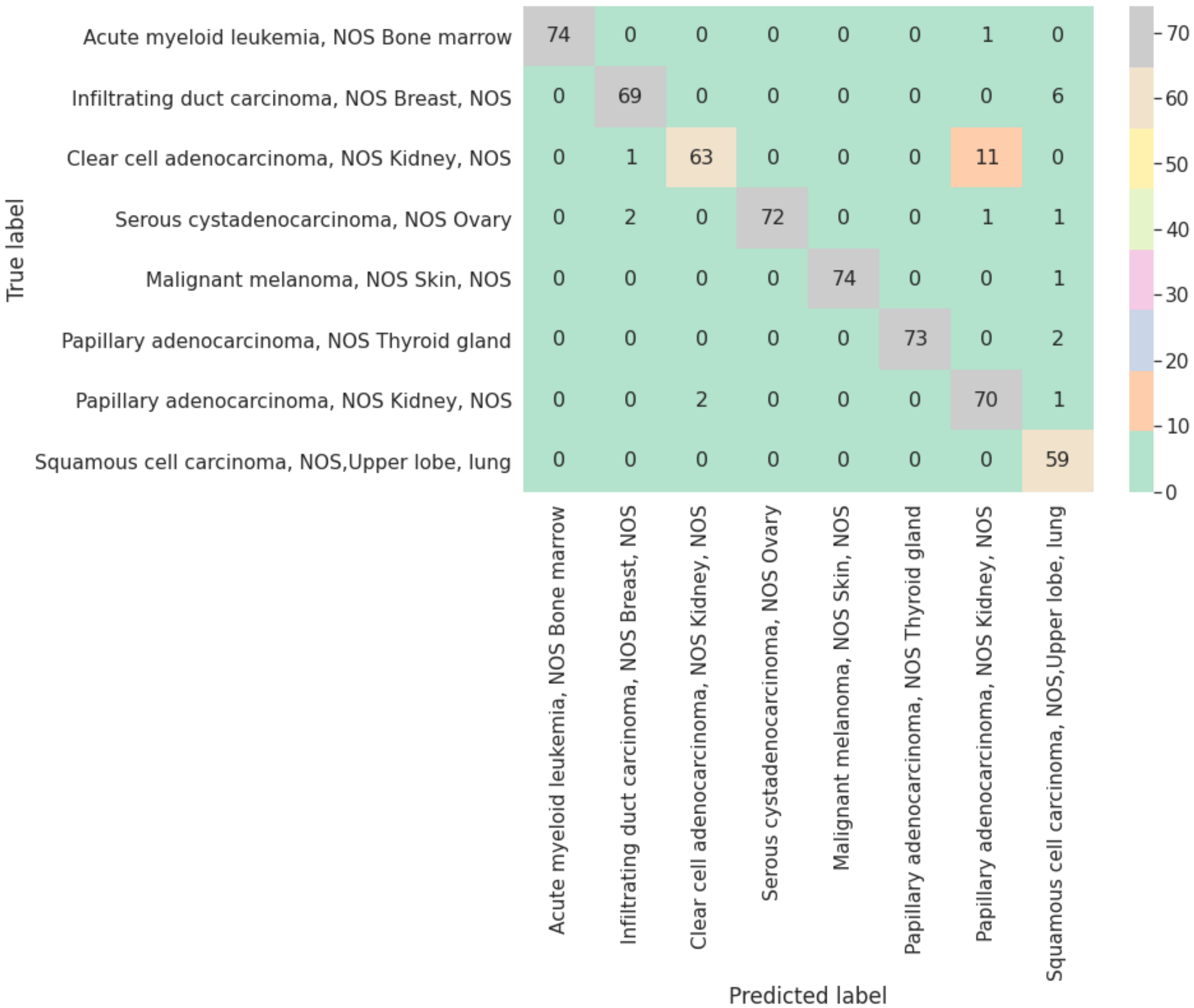
	Acute myeloid leukemia, NOS Bone marrow	Infiltrating duct carcinoma, NOS Breast, NOS	Clear cell adenocarcinoma, NOS Kidney, NOS	Serous cystadenocarcinoma, NOS Ovary	Malignant melanoma, NOS Skin, NOS	Papillary adenocarcinoma, NOS Thyroid gland	Papillary adenocarcinoma, NOS Kidney, NOS	Squamous cell carcinoma, NOS,Upper lobe, lung
Acute myeloid leukemia, NOS Bone marrow	75	0	0	0	0	0	0	0
Infiltrating duct carcinoma, NOS Breast, NOS	0	70	0	0	1	0	0	4
Clear cell adenocarcinoma, NOS Kidney, NOS	0	1	73	0	0	0	1	0
Serous cystadenocarcinoma, NOS Ovary	0	1	0	73	1	0	0	1
Malignant melanoma, NOS Skin, NOS	1	0	0	0	74	0	0	0
Papillary adenocarcinoma, NOS Thyroid gland	0	0	0	0	0	75	0	0
Papillary adenocarcinoma, NOS Kidney, NOS	0	0	4	0	0	0	68	1
Squamous cell carcinoma, NOS,Upper lobe, lung	0	1	0	0	0	0	0	58

Figure 20: Confusion Matrix of Artificial Neural Network



	Acute myeloid leukemia, NOS Bone marrow	Infiltrating duct carcinoma, NOS Breast, NOS	Clear cell adenocarcinoma, NOS Kidney, NOS	Serous cystadenocarcinoma, NOS Ovary	Malignant melanoma, NOS Skin, NOS	Papillary adenocarcinoma, NOS Thyroid gland	Papillary adenocarcinoma, NOS Kidney, NOS	Squamous cell carcinoma, NOS,Upper lobe, lung
Acute myeloid leukemia, NOS Bone marrow	75	0	0	0	0	0	0	0
Infiltrating duct carcinoma, NOS Breast, NOS	0	68	0	0	0	2	0	5
Clear cell adenocarcinoma, NOS Kidney, NOS	0	0	73	0	0	0	1	1
Serous cystadenocarcinoma, NOS Ovary	0	2	0	74	0	0	0	0
Malignant melanoma, NOS Skin, NOS	1	1	0	0	73	0	0	0
Papillary adenocarcinoma, NOS Thyroid gland	0	0	0	0	0	75	0	0
Papillary adenocarcinoma, NOS Kidney, NOS	0	1	3	0	0	0	69	0
Squamous cell carcinoma, NOS,Upper lobe, lung	0	1	0	0	0	0	0	58

Figure 21: Confusion Matrix of CResNet



True label \ Predicted label	Acute myeloid leukemia, NOS Bone marrow	Infiltrating duct carcinoma, NOS Breast, NOS	Clear cell adenocarcinoma, NOS Kidney, NOS	Serous cystadenocarcinoma, NOS Ovary	Malignant melanoma, NOS Skin, NOS	Papillary adenocarcinoma, NOS Thyroid gland	Papillary adenocarcinoma, NOS Kidney, NOS	Squamous cell carcinoma, NOS,Upper lobe, lung
Acute myeloid leukemia, NOS Bone marrow	74	0	0	0	0	0	1	0
Infiltrating duct carcinoma, NOS Breast, NOS	0	69	0	0	0	0	0	6
Clear cell adenocarcinoma, NOS Kidney, NOS	0	1	63	0	0	0	11	0
Serous cystadenocarcinoma, NOS Ovary	0	2	0	72	0	0	1	1
Malignant melanoma, NOS Skin, NOS	0	0	0	0	74	0	0	1
Papillary adenocarcinoma, NOS Thyroid gland	0	0	0	0	0	73	0	2
Papillary adenocarcinoma, NOS Kidney, NOS	0	0	2	0	0	0	70	1
Squamous cell carcinoma, NOS,Upper lobe, lung	0	0	0	0	0	0	0	59

Figure 22: Confusion Matrix of Ensembles

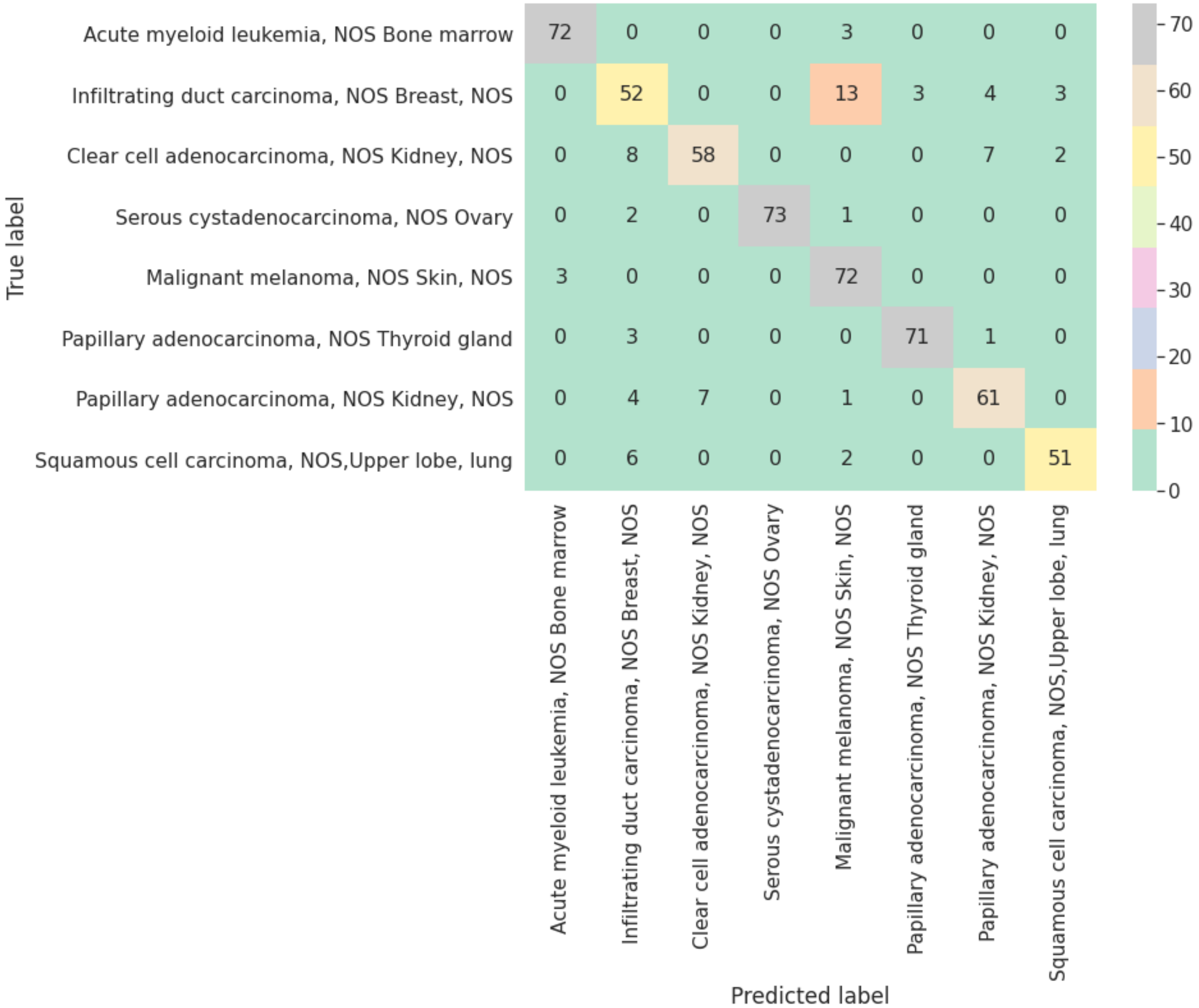


Figure 23: Confusion Matrix of LSTM

5.3 Mathew's Correlation Coefficient

In order to find out how the models have performed with existing architectures, it was imperative to calculate Mathew's Correlation Coefficient. In order to do that the we needed the model's prediction on test data which will give us true positives, true negatives, false positive and false negative. These are defined below:

- **True positive :** A value in a confusion matrix is true positive if the predicted label and the actual label are positive or true
- **True negative :** A value in a confusion matrix is true negative if the predicted label and the actual label are negative. Which means that the class is not present in this set of features and this absence of class has been picked up by the model
- **False positive :** A value in a confusion matrix is false positive if the predicted label is true and the actual label is false. This means despite the absence of the class in the data, the model gives a false prediction about the presence of the class in the data
- **False Negative :** A value in a confusion matrix is false negative if the predicted label is false and the actual label is true. This means that if the class is present in the data the model fails to identify that class given the data

For our cancer classification task it was important that the models gave low number of false negatives. This is because if a patient has cancer and the model fails to detect cancer, then the person would end up suffering due to mis diagnosis. Fortunately, our models gave a high MCC score and our CResNet model and ANN model has beaten the LSTM model in terms of MCC which can be observed from the tables. The MCC will be calculated using the equation 15

The tables 3, 4, 5, 6 represents the calculation of true positives, true negatives, false positives and false negatives of our ANN, CResNet, Ensemble and LSTM model respectively

6 Future Work and Conclusion

In conclusion, we believe that our approach to classifying cancer sub types will be less invasive and less expensive and has the potential to save large amounts of life through early diagnosis.We have introduced 4 types of models ANN, CResNet, Ensemble and LSTM . We saw that the best model which gave us the best results was using the CResNet model and the second best result was given by our ANN model. Our ANN model and CResNet model gave an MCC score of 0.9624625 and 0.97445 respectively. This is better than the architecture LSTM which gave an MCC value of 0.859625.

For future works the interpretability between the features and classification can be explored. So that when a patient is diagnosed with cancer the doctor or clinician will be able to immediately recognize which of the 1881 features are responsible for this particular type of cancer. This will offer the doctors an easier diagnosis and make the treatment of a cancer patients much smoother.

ANN					
Disease	True Positive	True Negative	False Positive	False Negative	MCC (Matthew's Correlation Coefficient
Acute myeloid leukemia, NOS Bone marrow	75	507	1	0	0.9924
Infiltrating duct carcinoma, NOS Breast, NOS	70	505	3	5	0.9084
Clear cell adenocarcinoma, NOS Kidney, NOS	73	505	4	2	0.9547
Serous cystadenocarcinoma, NOS Ovary	73	508	0	3	0.9772
Malignant melanoma, NOS Skin, NOS	74	509	2	1	0.9772
Papillary adenocarcinoma, NOS Thyroid gland	75	508	0	0	1
Papillary adenocarcinoma , NOS Kidney, NOS	68	514	1	5	0.9524
Squamous cell carcinoma, NOS Upper lobe, Lung	58	525	6	1	0.9374
Average					0.9624625

Table 3: MCC for ANN

CResNet					
Disease	True Positive	True Negative	False Positive	False Negative	MCC (Matthew's Correlation Coefficient
Acute myeloid leukemia, NOS Bone marrow	75	507	1	0	0.9924
Infiltrating duct carcinoma, NOS Breast, NOS	71	505	3	4	0.9462
Clear cell adenocarcinoma, NOS Kidney, NOS	73	506	2	2	0.9694
Serous cystadenocarcinoma, NOS Ovary	74	506	1	2	0.9772
Malignant melanoma, NOS Skin, NOS	73	507	1	2	0.9769
Papillary adenocarcinoma, NOS Thyroid gland	75	507	1	0	0.9924
Papillary adenocarcinoma , NOS Kidney, NOS	70	510	1	3	0.9684
Squamous cell carcinoma, NOS Upper lobe, Lung	59	521	3	0	0.9727
Average					0.97445

Table 4: MCC for CResNet

Ensemble					
Disease	True Positive	True Negative	False Positive	False Negative	MCC (Matthew's Correlation Coefficient
Acute myeloid leukemia, NOS Bone marrow	74	508	0	1	0.9923
Infiltrating duct carcinoma, NOS Breast, NOS	69	506	3	6	0.9302
Clear cell adenocarcinoma, NOS Kidney, NOS	63	508	2	12	0.8894
Serous cystadenocarcinoma, NOS Ovary	72	510	0	4	0.9695
Malignant melanoma, NOS Skin, NOS	74	509	0	1	0.9923
Papillary adenocarcinoma, NOS Thyroid gland	73	509	0	2	0.9846
Papillary adenocarcinoma , NOS Kidney, NOS	70	513	13	3	0.8846
Squamous cell carcinoma, NOS Upper lobe, Lung	59	525	11	0	0.9086
Average					0.9439375

Table 5: MCC of Ensembles

LSTM					
Disease	True Positive	True Negative	False Positive	False Negative	MCC (Matthew's Correlation Coefficient
Acute myeloid leukemia, NOS Bone marrow	72	505	3	3	0.9541
Infiltrating duct carcinoma, NOS Breast, NOS	52	485	23	23	0.6481
Clear cell adenocarcinoma, NOS Kidney, NOS	58	509	7	17	0.8083
Serous cystadenocarcinoma, NOS Ovary	73	509	0	3	0.9772
Malignant melanoma, NOS Skin, NOS	72	508	20	3	0.8463
Papillary adenocarcinoma, NOS Thyroid gland	71	511	3	4	0.9462
Papillary adenocarcinoma , NOS Kidney, NOS	61	522	12	12	0.8131
Squamous cell carcinoma, NOS Upper lobe, Lung	51	532	5	7	0.8837
Average					0.859625

Table 6: MCC of LSTM

References

[1] Claudio R Alarcón et al. "N 6-methyladenosine marks primary microRNAs for processing". In: *Nature* 519.7544 (2015), pp. 482–485.

[2] George Adrian Calin et al. "Frequent deletions and down-regulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia". In: *Proceedings of the national academy of sciences* 99.24 (2002), pp. 15524–15529.

[3] Jordan M Cummins et al. "The colorectal microRNAome". In: *Proceedings of the National Academy of Sciences* 103.10 (2006), pp. 3687–3692.

[4] Terrance DeVries and Graham W Taylor. "Dataset augmentation in feature space". In: *arXiv preprint arXiv:1702.05538* (2017).

[5] E Melo Felipe De Sousa et al. "Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions". In: *Nature medicine* 19.5 (2013), pp. 614–618.

[6] Douglas Hanahan and Robert A Weinberg. "Hallmarks of cancer: the next generation". In: *cell* 144.5 (2011), pp. 646–674.

[7] Yoji Hayashita et al. "A polycistronic microRNA cluster, miR-17-92, is overexpressed in human lung cancers and enhances cell proliferation". In: *Cancer research* 65.21 (2005), pp. 9628–9632.

[8] JR Jass. "Classification of colorectal cancer based on correlation of clinical, morphological and molecular features". In: *Histopathology* 50.1 (2007), pp. 113–130.

[9] Jun Lu et al. "MicroRNA expression profiles classify human cancers". In: *nature* 435.7043 (2005), pp. 834–838.

[10] Xiaoya Luo et al. "MicroRNA signatures: novel biomarker for colorectal cancer?" In: *Cancer Epidemiology and Prevention Biomarkers* 20.7 (2011), pp. 1272–1286.

[11] Brian W Matthews. "Comparison of the predicted and observed secondary structure of T4 phage lysozyme". In: *Biochimica et Biophysica Acta (BBA)-Protein Structure* 405.2 (1975), pp. 442–451.

[12] Konstantinos J Mavrakis et al. "Genome-wide RNA-mediated interference screen identifies miR-19 targets in Notch-induced T-cell acute lymphoblastic leukaemia". In: *Nature cell biology* 12.4 (2010), pp. 372–379.

[13] Ali Muhamed Ali et al. "A machine learning approach for the classification of kidney cancer subtypes using mirna genome data". In: *Applied Sciences* 8.12 (2018), p. 2422.

[14] Ann L Oberg et al. "miRNA expression in colon polyps provides evidence for a multihit model of colon cancer". In: *PloS one* 6.6 (2011), e20465.

[15] Yong Peng and Carlo M Croce. "The role of MicroRNAs in human cancer". In: *Signal transduction and targeted therapy* 1.1 (2016), pp. 1–9.

[16] Aaron J Schetter et al. "MicroRNA expression profiles associated with prognosis and therapeutic outcome in colon adenocarcinoma". In: *Jama* 299.4 (2008), pp. 425–436.

[17] Aaron J Schetter, Hirokazu Okayama, and Curtis C Harris. "The role of microRNAs in colorectal cancer". In: *Cancer journal (Sudbury, Mass.)* 18.3 (2012), p. 244.

[18] Mary Shapcott, Katherine J Hewitt, and Nasir Rajpoot. "Deep Learning With Sampling in Colon Cancer Histology". In: *Frontiers in Bioengineering and Biotechnology* 7 (2019), p. 52.

[19] Yingshuai Sun et al. “Identification of 12 cancer types through genome deep learning”. In: *Scientific reports* 9.1 (2019), pp. 1–9.

[20] H Tagawa and M Seto. “A microRNA cluster as a target of genomic amplification in malignant lymphoma”. In: *Leukemia* 19.11 (2005), pp. 2013–2016.

[21] Justin L Wang et al. “Classification of white blood cells with patternnet-fused ensemble of convolutional neural networks (pecnn)”. In: *2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*. IEEE. 2018, pp. 325–330.

[22] Wei Yang, Kuanquan Wang, and Wangmeng Zuo. “Neighborhood Component Feature Selection for High-Dimensional Data.” In: *JCP* 7.1 (2012), pp. 161–168.

[23] Lin Zhang et al. “microRNAs exhibit high frequency genomic alterations in human cancer”. In: *Proceedings of the National Academy of Sciences* 103.24 (2006), pp. 9136–9141.