Cancer Classification Based on Genetic Data Using Stacked Autoencoder-DNN Model

A THESIS SUBMITTED

FOR THE DEGREE OF

MASTER OF TECHNOLOGY

IN

INFORMATION TECHNOLOGY

by

DEEPALI JINDAL

under the guidance of

Dr. MOHAMMED JAVED

Department of Information Technology

Indian Institute of Information Technology-Allahabad

July 2021

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY ALLAHABAD

(A Centre of Excellence in Information Technology Established by Govt. of India)

CANDIDATE DECLARATION

I, your name, Roll no: MIT2019006 certify that this thesis entitled, 'Cancer Classification based on genetic data using Stacked Autoencoder-DNN Model' is submitted by me in partial fulfillment of the requirement of the Degree of Master of Technology in Information Technology, Indian Institute of Information Technology, Allahabad.I understand that plagiarism includes:

- Reproducing someone else's work (fully or partially) or ideas and claiming it as one's own.
- Reproducing someone else's work (Verbatim copying or paraphrasing) without crediting.
- Committing literary theft (copying some unique literary construct).

I have given due credit to the original authors/ sources through proper citation for all the words, ideas, diagrams, graphics, computer programs, experiments, results, websites, that are not my original contribution. I have used quotation marks to identify verbatim sentences and given credit to the original authors/sources.
I affirm that no portion of my work is plagiarized. In the event of a complaint of plagiarism, I shall be fully responsible. I understand that my Supervisor may not be in a position to verify that this work is not plagiarized.
Name: **Deepali Jindal**
Enrolment No: **MIT2019006**
Dept of: **Information Technology**

Signature: Deepali Jindal

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY

ALLAHABAD

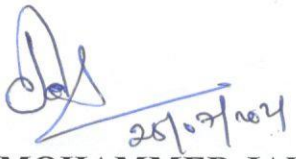
(A Centre of Excellence in Information Technology Established by Govt. of India)

CERTIFICATE FROM SUPERVISOR

This is to certify that the statement made by the candidate is correct to the best of my knowledge and belief. The master's thesis titled "**Cancer Classification based on genetic data using Stacked Autoencoder-DNN Model**" by **Deepali Jindal (MIT2019006),** has been carried out under my supervision and guidance and that this work has not been submitted elsewhere for a degree. I do hereby recommend that it should be accepted in the fulfillment of the requirements of the Master's Thesis at IIIT Allahabad.

Date: / /

Place: Allahabad



(Dr. MOHAMMED JAVED)

Counter Signed by Dean (Academics)

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY

ALLAHABAD

(A Centre of Excellence in Information Technology Established by Govt. of India)

CERTIFICATE OF APPROVAL

The forgoing thesis titled "**Cancer Classification based on genetic data using Stacked Autoencoder-DNN Model**" by **Deepali Jindal (MIT2019006)** is hereby approved as a credible study in the field of Information Technology carried out and presented in a manner satisfactory to warrant its acceptance as a prerequisite to the degree for which it has been submitted. It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made opinion expressed or conclusion drawn therein but approve the the thesis only for the purpose for which it is submitted.

Dr. Mohammad Javed, Supervisor, IIITA

Dr. Navjot Singh, HOD-IT Nominee, IIITA

Dr. Ramesh K. Bhukya, Expert, IIITA

Dr. Bunil Balabantaray, External Expert
NIT MEGHALAYA

Signature of the Examiners Committee Members

(On final examination and approval of the thesis)

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY ALLAHABAD

(A Centre of Excellence in Information Technology Established by Govt. of India)

KEYWORDS

Cancer Classification

* Genetic Mutation

** Gene Expression Data

Deep Learning

* Deep Neural Network

** Autoencoder

ABSTRACT

Cancer is one of the prime causes for deaths all over the world with about 10 million deaths in 2020 alone. The sequence of genes gets disrupted in human's body which leads to a genetic mutation and eventually leads to growth of tumor or cancerous cells. An early detection can reduce the total number of deaths by a significant margin. With the advancements of modern artificial intelligence detecting the early mutation can be revolutionized and can be done even without doctor's aid or examining the patient. Cancer detection is one of the promising areas of research and emerging as a center of focus for the healthcare industry. Autoencoder is a feature transformation technique which works better than traditional techniques like PCA when the relationship between the features in the data is nonlinear. In recent years, several deep learning-based classification approaches have been developed in the context of cancer detection specifically based on the concept of Machine Learning. Concerning the same, the present thesis proposes a usage of a stacking technique using two deep learning models i.e Autoencoders and Deep neural network. The performance of proposed architecture is further compared with existing architecture. Experimental results show that the proposed stacked autoencoder with deep neural network architecture has outperformed the discussed approach in terms of accuracy.

Contents

List of Tables

List of Figures

Acknowledgements

The satisfaction that accompanies the successful completion of any project work would be impossible without the mention of the people who made it possible and whose constant guidance and encouragement crown all the efforts. This thesis was not only an endeavor but also an interesting learning experience for me and it bears the imprint of a number of people who directly or indirectly were a great source of help and constant encouragement.

I would like to express my sincere thanks to my mentor Dr. Mohammed Javed for his continuous motivation and guidance. Their valuable suggestions, comments and support were an immense help for me. I am grateful to them for taking out time from their busy schedule and being very supportive in guiding my work.

I would also like to thank the research scholars Mr. Bulla Rajesh for helping and guiding me throughout with the technical issues. Special thanks to my parents, my guardians and friends at IIIT-A for their constant support and motivation. The interesting and informative discussions we had together greatly contributes to the completion of this work.

Deepali Jindal

Chapter 1

Introduction

1.1 CANCER: Bane to the world

1.1.1 Overview

Since it was known to exist, Cancer has been the most evil foe of life. It is deadly, complex and humans are still struggling to find one cure for all types of Cancer [7].

Human body is made of trillions of smaller living units called cells, and Cancer can get started in any part of the body[8]. Cancer is a disease which happens because of uncontrollable growth of cells that can arise in any type of cell of the body. Those uncontrollable cells can break and will get spread to other parts of the body.

Generally, the cells in the human body grow through a process of cell division. Once the damaged cells die, they get substituted by new cells. When this structured system gets disturbed, the rapid growth of abnormal cells accelerates and spreads to other organs, when they should not. These abnormal cells lead to growth of tumors, and this process is known as Metastases. Metastases are the reason for death from cancer in most of the cases [7, 9].

Cancer has been with us since life has been with us. The reason why Cancer is so deadly is, it is not a foreign virus or microbe, it can trigger in anyone's body at any point of time.

It is a very complex disease. That is one of the reasons why it is such an enigma. There are multiple external causes of cancer such as smoking, carcinogens, viruses, and excessive drinking. Although, the important internal cause of the disease is purely genetic.

Some cancers like retinoblastoma are relatively simple, while others like breast, lungs, etc are complex genetic aberrations that are difficult to pinpoint. Multiple mutations in the genetic code result in cancer [10]. The body does not function the same all the time. As we grow old, the capability of the body to eliminate those cells which have damaged DNA reduces in comparison to the early stage of life. And hence, it is one of the reasons to have more chances of cancer at later stages of a person's life.

1.1.2 How Cancer grows and impact on different parts of body?

Cancer can get started in any organ of the body. So, depending on the organ where the tumor arises, the type/name of the cancer is derived. Some of the types of cancer are eye cancer, lung cancer, bladder cancer and many more. Typically, we can categorize more than 200 cancer cells in 5 categories, that are described below [11].

- Carcinoma : These are the most common types of cancer cells. The origin of these cells in the body is epithelia. Epithelia tissue is a collection of cells which are tightly packed together having very little space between their membranes. It lines internal closed cavities as well as covers external surfaces of the body [11, 12]. Carcinomas are various types of epithelia cells that can originate in different carcinomas, that are Adenocarcinoma, Squamous, Basal and Transitional cell carcinoma.

- Sarcoma : Sarcomas originate in connective tissues. These are also considered as supporting tissues in the human body. The cancers formed in bone. Soft tissues, blood vessels, fat and lymph vessels are Sarcomas cancer. This accounts for one in every hundred cancers which gets diagnosed every single year. These are more common than carcinoma and can be categorized in two types : Bone Sarcomas and Soft Tissue Sarcomas.

- Leukemia : Cancer which occurs in those tissues which helps in blood formation and causes a blood disorder is known as leukemia. When bone marrow makes too many WBCs, they crowd out normal blood cells. Leukemia does not form any tumor but an abnormal number of white blood cells. Due to overgrowth of bad cells over normal cells, the body finds it difficult to provide sufficient oxygen to its tissues, and also face problems in controlling the bleeding or fighting back with infections [13]. Leukemias are rare. It makes up to 3 percent of all cancer cases. But, these are found commonly in children.

- Lymphoma and myeloma : Lymphoma cancer is formed with swollen lymph nodes where lymphocytes mature in the body. These are crucial parts of the immune system which fights the diseases. Due to the formation of these damaged lymphocytes in lymph nodes, it can get started from anywhere.

  Multiple myeloma involves a certain cell called a plasma cell. This particular cell which is part of the immune system goes from good to bad which can be seen by observing elevation in protein levels. It can spread all through the body by forming multiple tumors in the bones [11, 14].

- Brain and Spinal cord cancers : Brain cancers grow from cells within the brain that support the normal neurons of the brain. These cells can lose their genetic ability to control their normal growth and start to grow in an uncontrolled fashion which causes the tumor. Some of the tumors spread rapidly and some spread slowly in the brain.

1.1.3 Main reason behind Cancer : Genetic Mutation

Cancer is known as a genetic disease which implies cancer is formed because of disrupted order in the sequence of genes which leads to loss control on normal mechanisms of the cell . Cancer starts when genes in a cell start to grow abnormally [7, 15].

Cancer can be transmitted through parents to children if the genetic changes occurred in reproductive cells which are germs cells of the body. These changes are known as germline changes which are present in every cell of the children.

Mutations can be explained as abnormal changes in the DNA of a gene. These involve changes in the arrangement of the bases that make up a gene [15, 16]. A gene mutation can affect the cell in various ways such as - some mutations stop the production of protein, others may change the protein in a way that it works differently or doesn't work at all.

In a few cases, mutations may cause a gene to be turned on, and make more of the protein than required. Some mutations don't have any noticeable effect, but others may invite a life-threatening disease. In most of the cases, many turns of mutations are needed before a cell can become a cancer cell. The mutations can affect genes that control the mechanism of cell growth. Some can be called tumor suppressor genes. Often, mutations may also lead normal genes to become cancer-causing genes which are widely known as oncogenes.

Every cancer is caused by mutations in genes. In figure 1.1, it is shown how normal cells turn into tumorous cells due to the harmful mutations that take place. Mutations can be the result of several factors which include aging, environmental conditions, toxic radiations, smoking, certain chemicals like nitrate and infectious agents [16–18]. A large number of mutations may happen in one cell with the time, which may turn into cancerous cells.This process takes a long time, and this can be a valid explanation why in certain people cancer occurs at a later stage in life. If the mutation has happened because of a certain external factor, they cannot pass them on to their children.

On the other hand, a very high number of cancer cases are considered as inheritable which passed down to offspring. If the genetic mutation happens in the

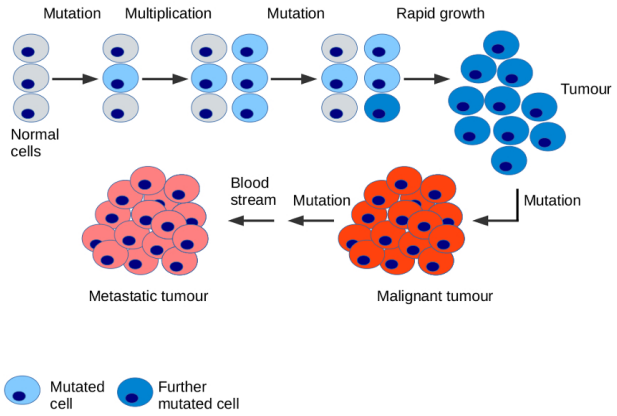


FIGURE 1.1: Evolution of cancer in cells [1]

reproductive cell and passed down to offspring, then in this case, the mutation will be present in each and every cell of the newborn. This is the reason why in some families, members are expected to have certain kinds of cancer.

1.1.3.1 Different types of Genetic Mutation

- Point mutation : In this type of mutation, there is change only in the single base of the base of the DNA. It is commonly known as single nucleotide polymorphism.
- Insertion Mutation : In this type of mutation, there is an insertion or addition of a base into the genetic sequence. This is commonly known as addition mutation.
- Deletion Mutation : When there is a deletion of single or multiple bases in a genetic sequence, it is known as deletion mutation.
- Inversion : When there are a few genetic sequences inverted or inserted back into the original sequence, it is referred as Inversion. Substitution– When a few bases of a genetic sequence are replaced by other bases, it is called substitution.
- Duplication : When a single base or multiple bases occur multiple times in a genetic sequence, it is known as duplication.

1.1.3.2 Cancer causing Genes

There are various types of mutated genes that can cause cancer which has been shortly described below. The reason behind the repeated occurrence of mutation still remains unknown to a greater extent. Mutations can be caused by certain external factors (discussed previously) in the environment. However, the progress of continuous mutation is a native part of aging.

- Oncogenes: There are genes that code for proteins that normally direct cell growth and start out as proto-oncogenes. When altered or mutated, the proto-oncogene transforms into a full blown oncogene [19]. Once turned to oncogene, they escalate the growth of tumor. Some of the properties of oncogene are listed below:
  - These are acquired mutations.
  - These are known to be primarily dominant at the plasmic level. Even if the mutation is in one pair of proto-oncogene only, it can solely become a reason to make changes in the evolution of cells and the growth of a tumor.
  - Genes which promote tumor growth are known to facilitate the formation of tumors or to segregate specific factors of progression tumor.
- Tumor suppressor genes: The genes which suppress tumors are generally known to be found in our cell. In a healthy state, they help to regulate processes consisting of cell growth and keep a check on death of cells. It directly helps to suppress the tumor growth. When such gene gets altered, it causes tumor growth [19]. Some of the properties of tumor suppressor genes are listed below:
  - Both the replicas of gene pairs are mutated to lead to a change in cell growth of cells. The sole reason why suppressor genes act recessively at the plasmic level is this only.
  - Most of the times mutations occur in tumor suppressor genes are acquired. Mutations might occur due to prolonged aging, external environmental aspects or may be both.

  - In a few cases, the mutation can be inherited in a tumor suppressor gene.
  - In case of heredity cancer, tumor suppressor genes are found in the majority. But these mutations are not necessarily inherited.

- The DNA in a cell makes a replica of itself through cell division [20]. During this complex process, errors can happen. If this error has happened in tumor suppressor genes or what we call as proto-oncogenes, it can become the cause of abnormal cell growth. Some of the properties of DNA repair genes are listed below:

  - They can be an inherent mutant gene.
  - They can be acquired over a period of time as the outcome of aging and external aspects.
  - It is necessary that both the copies of the gene-pair of DNA repair genes are mutated. This is why it is known to be recessive at plasmic level.

1.1.4 Challenges in Cancer Diagnosis

Treating cancer is undoubtedly a herculean task. As published by the department of oncology, the survival rate of a person is just 14% if it is detected at stage 4, since detecting the disease at early stages is arduous because of not having more symptoms [21]. At later stages treatment is only about controlling the disease as long as they can.

The important process in the detection of tumors is collecting the sample of tissues. But this is not as easy as it looks, as it is quite difficult to decide whether to take tissue samples or not. The reason behind this is cancer spreads rapidly which makes it difficult to control.

Biopsies is another choice for detection of cancer, but due to its complex process it becomes difficult to take for some cancers. Lung collapse and other issues like bleeding are serious threats associated while taking biopsies in account [21, 22]. Another sample has to be collected if a particular sample is wrong or gives less

knowledge about the disease. The lack of knowledge and information about the disease may lead to the wrong diagnosis for the patient.

One of the game changing diagnoses of cancer is gene sequencing which may help to identify genetic variants that happened by just looking at the same cancer sample [16]. However, this process has its own hurdles like it is complicated to handle such a huge amount of data, as this generates terabytes of data [21]. And not all genetic variants are part of the growth of tumors, some of the gene variants that lead to tumor growth.

1.1.4.1 Need of Early Detection

Deaths due to cancer can be considerably reduced if cases are diagnosed and treatment is started at an early stage [23]. If diagnosed early, there is a high chance that cancer will respond positively to the undergone treatment which can result in a high survival rate and low death rate. This will also cut down the expenses incurred in the expensive treatment. Without any doubt, prominent progress can be resulted in the lives of cancer survivors by early detection [22, 23]. Spreading awareness about the probable and known symptoms of different types of cancer and the significance of going for medical advice plays an important role in early diagnosis. Also easy access to clinical evaluation and diagnostic services is important. Early treatment of symptomatic cancers is significant in all varied types of cancers. Cancer programs should be effectively designed to eliminate reduction in delays, remove pit holes on the way of diagnosis, expedite the treatment and curate cancer care initiatives.

1.2 Literature Review

1.2.1 Existing Work

To classify different types of tumor plays a significant role in curing cancer and medication. However, a great number of detection of cancer done previously are clinical based and have little reach to curing ability. Use of genetic data to classify

the types of cancer is known to have the potential to address the core problems related to curing cancer and medication [24, 25].

The introduction of DNA microarray technique has enabled us to monitor hundreds-thousands expressions of genes. One of the most disruptive discoveries in cancer diagnosis is - Microarrays [26]. It enables us to simultaneously monitor ten thousand expressions of genes and is amounting to informative data.

The analysis and management of such data is a huge obstacle between fully leveraging the capabilities of this technique. The unstructured data of microarray is images that are further changed into expressional gene matrices, where genes are represented in rows and columns are represented by varied samples of tissues or external factors [27]. The number in a particular cell characterizes the expression level of different gene in the distinct samples.

A further analysis has to be performed in order to extract in depth knowledge of different genetic processes. With such a huge data of gene expression, researchers are trying to look into probable cancer classification leveraging this available data. A considerable number of methods have been put forward with potential results which have been discussed below.

Joseph M. De Guia has presented his work on the classification of cancer using gene expression data [28]. The paper describes the gene selection as an important phase during classification tasks. The dimensional reduction technique using PCA has been used for gene selection before feeding it to the classification model [29]. Dimensionality reduction is a technique used to reduce the number of features in the dataset to fasten the learning process for deep learning and machine learning algorithms. It will give better training time and is also capable of storing it easily as the size of the dataset gets reduced after removing the unuseful features from the dataset. Large number of features can adversely affect the quality of the machine learning models fit on the data also referred to as the ‘curse of dimensionality. The paper has used the SVM and Boosting algorithm as a classifier where the accuracy obtained by SVM is 58% and 64% by XGBoost. Further research has been done on gene expression data using deep learning methods and shows better performance.

Omar Ahmed has published the research work on cancer classification based on gene expression data where they used four different deep Learning algorithms which includes Deep Neural Network (DNN), improved DNN by adding 20% droupout in the architecture of DNN, Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) [3]. The models have been tested and trained on four different benchmark datasets which includes DLBCL, Prostate Tumor, Leukemia and Colon Tumor. Improved DNN has shown the best results among all the other algorithms on these four datasets. For DLBCL, improved DNN gave 98.4% accuracy, CNN and RNN gave 96.8% accuracy. However, the results do not provide good results on Prostate and Colon tumors dataset as compared to Leukemia and DLBCL dataset.

1.2.2 Classification Algorithms

1.2.2.1 Support Vectore Machine

This is one of the most popular classification techniques in supervised learning algorithms. SVM can also be used for regression problems however the use of SVM is more suitable for classification problems [30]. The algorithm divides n-dimensional space into different classes by creating the best line or decision boundary. This best boundary is also called the best decision boundary. SVM creates hyperplanes by selecting the corner points/vectors and these corner vectors are called support vectors.

Types of SVM

- Linear SVM : The data that can be separated into two categories by drawing a straight line is called linearly separable data and we can use Linear SVM to solve such problems.
- Non-linear SVM : The data that cannot be separated into two categories by drawing a straight line is called non-linear data and we can use Non-Linear SVM to solve such problems.

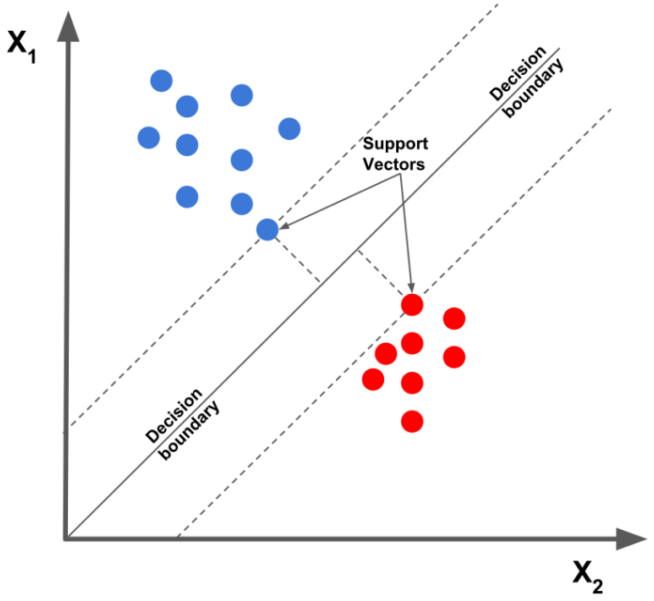


FIGURE 1.2: SVM with three hyper planes[2]

Hyperplane : Many decision boundaries can be drawn to divide the n-dimensional space into different target categories, however SVM will find the best decision boundary that will categorise the data. The dimensions of the hyperplane depend on the unique number of categorical values in the target variable.

Support vectors : Support vectors are those vectors that are closest to the hyperplane which impact the position of the hyperplane are called Support Vector which has been shown in figure 1.2. Since these support the hyperplane, these are called support vectors.

1.2.2.2 Extreme Gradient Boosting

XGBoost is a decision tree based machine learning algorithm which uses gradient boost framework to implement it [31]. It works well with small to medium datasets and is used for classification and regression problems. The iterative approach has been used by boosting which is used to build a tree, then compute the result of an objective function and then feed the result which is obtained as an input to the following tree's objective function. The architecture of XGBoost has been shown in figure 1.3. The output produced by the previous tree's model will be given as an

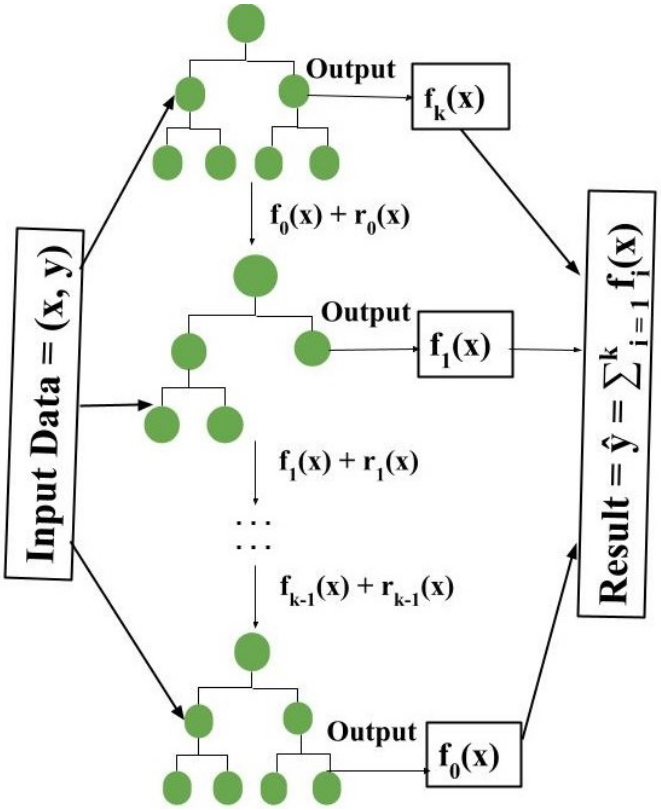


FIGURE 1.3: Architecture of XG Boost

input to the next tree's model, after improving its mistakes, which lead to overall improvement in results [32].

Boosting modifies the weights affiliated with every single split of a tree on the basis of error obtained by the previous ones. Gradient Boosting algorithm, optimizes your objective function.

1.2.2.3 Naive Bayesian

A classification technique based on bayes theorem. Naive bayes makes two naive assumptions to make the classification. These assumptions are normally not correct but these assumptions allow us to work with naive bayes which works fairly well with these assumptions put in place [33].

- All the features are independent of the outcome.

- Equally contribute to the outcome.

Bayes theorem : This theorem provides the basis for Bayesian inference. It is used to compute the probability of an event (X) which is happening given the probability of an event (Y) which has previously happened. This theorem can be mathematically stated as:

$$P(X|Y) \ = \ \frac{P(Y|X)P(X)}{P(Y)}$$

(1.1)

Where X represents that event which will occur and Y represents that event which has already occurred.

P(X) is priori of X

P(Y) denotes the posterior probability of Y.

Naive bayes is very fast compared to many algorithms because closed form expression calculation helps in faster training of maximum likelihood as it grows at linear rate versus other iterative approximations used for other classifiers.

1.2.2.4 Convolutional Neural Network

The CNN algorithm has huge potential in categorizing patterns that can be leveraged to generate further eccentric patterns within advancing layers [34]. This is an outstanding classification of multi-layered neural networks. Compared to other neural networks it also leverages the technique of back propagation algorithm for training purposes.

CNN's architecture is a key differentiator among others. The general architecture of CNN has been shown in figure 1.4. Generally, it has three layers: the first layer is an input layer, the next layer is a hidden layer which can have multiple hidden layers as a part of the hidden layer and the last layer is the output layer.

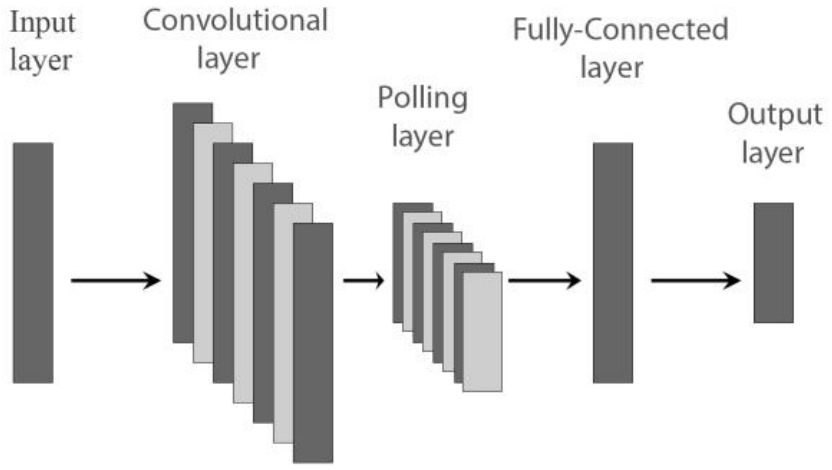


FIGURE 1.4: General Architecture of CNN [3]

Further, the multiple hidden layers are made up of several layers in which one of them is convolutional layers, the other one is pooling layers, and the last one is fully connected layers. Convolutional layer is the topmost layer, which is used to extort the varied features from the given data. And the computational operation of convolution is carried out between the input data and a filter of a specific size NxN. Pooling layer reduces the computational costs. It acts as a liaison between the Convolutional Layer and the Fully Connected Layer. And the purpose of a fully connected layer is to connect the underlying neurons between two layers. These layers are put right before the output layer and it forms the last few layers of the architecture.

1.2.2.5 Recurrent Neural Network

RNN is another neural network which mostly works on natural language processing tasks and also shows promising results in other domains as well [35]. In RNN, the input of the next layer is not only input data but also the previous output that it gave as shown in figure 1.5. Each node represents a context or memory which carries the information and learns from past events. Some of the usecases where RNN is used in our real life are text prediction in gmail, sentiment analysis like in amazon where rating is provided based on the reviews provided by customers, google translator and many more. RNN prefers sequence modelling tasks. Other

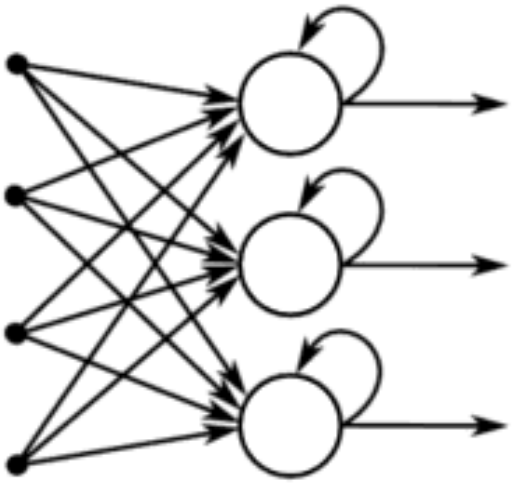


FIGURE 1.5: General Architecture of RNN [4]

neural networks may face computational issues and also no parameter sharing among them where sequence matters.

1.2.2.6 Deep Neural Network

DNN (Deep Neural Network) is inspired by the structure of neurons which are present in the brain of humans where each node in a neural network acts as a neuron which stimulates when any data or signal is passed to it. DNN results to be more complex in comparison to neural networks. It works more like a human which understands patterns from any data and tries to solve unseen problems by its own without having prior knowledge based on the data provided. However, neural networks work on a particular input data and solves problems based on the specific algorithms. It is more like machine learning where a model learns from input data and derives more valuable and complicated functions from a group of nodes which is represented as one layer. The DNN model has performed better when some of the nodes or connections have been dropped out randomly [36].

1.3 Problem Statement

Cancer is found to be one of the main reason of death worldwide, that accounts for more than 9 million deaths in 2020 alone. The cause of cancer is an unwanted mutation of genes in the cells. Changes may occur due to incorrect replication done by DNA or the outcome of environmental circumstances such as UV radiations, smoking etc. It is important and necessary in cancer study to detect and prognosis gene mutation which encourages tumor growth, as it will lead to precision medicine for cancer treatment. Researchers collect terabytes of information on gene mutation like gene expression data or clinical text data across the countries and use them in their studies to identify the complex pattern and extract useful information from collected genetic data. The process of determining genetic data can take weeks, or months or in some cases several years, which can result in extreme delay in treatment and care. That's why, there is a need to bring artificial intelligence into practice, as AI can extract meaningful relationships from raw data and can assist to classify cancer in a quick and efficient way.

For the purpose of this thesis we have taken gene expression data and applied deep learning techniques to perform classification. We have proposed a stacked neural networks approach to the gene classification problem by stacking autoencoder with a classical deep neural network. The purpose of the thesis is to empirically prove that meaningfully stacked models applied to the right problems like gene expression classification will perform better than the traditional single deep neural network. The input is gene expression data which has gene expression levels of different genes. The neural network is then fed with this expression data, post which it does a binary classification.

1.4 Proposed Methodology

A microarray is the technology which can track hundreds of genes at a time and help to track expression level of genes. This technology is helpful for detecting cancer classification based on its generation of data. The microarray got transferred into gene expression data which is used by researchers to detect cancer. In

this study, the gene expression data has been used to classify cancer. In order to get necessary genetic information from cancer related microarray data and to reduce high dimensionality problem, noise and understand the feature to classify the cancer, a stacked autoencoder DNN model has been proposed which is further extension of the DNN with dropout model.

1.5 Organizing the thesis

Chapter 1 laid down the biological background of cancer along with the challenges that occur in cancer diagnosis, the literature survey that has discussed the approach performed during the research and described the problem statement that motivated throughout to stick on this research.

Chapter 2 determines the dataset description and proposed methodology used to classify cancer classes.

Chapter 3 provides the details of software, libraries platforms exercised, and interpretation of results.

Chapter 4 concludes and analyzes the results with the future extension of the same.

1.6 Conclusion

In the present chapter, we learnt about the biological background of cancer, including what cancer is and how it develops. We dug in depth to understand the main reason for cancer to occur is genetic mutations. We studied how detection of genetic mutation which leads to tumor can help to diagnose cancer at an early stage.

We discussed elaborately how AI can fasten the process of predicting and detecting disease at an early stage. We briefly covered a few studies which were carried out on cancer classification using machine learning algorithms, based on genetic data and we discussed the scope of improvement in existing models.

In the upcoming chapter, we'll be discussing the proposed methodology in a detailed manner which will assist us to improve cancer classification based on genetic data.

Chapter 2

Cancer detection based on genetic data

2.1 Overview

Majority of treatments of patients with cancer rely on chemotherapy and other types of medication. However, these treatments don't provide uniformity in effectiveness, that's why it becomes important for the healthcare domain to have the good and suited drug treatment based on each individual. Digitized genetic data can be leveraged to greater extent in the field of getting right drug therapy.

As per some researchers, the better understanding of the molecular events at the origin of the disease or disorder can assist us in providing better insight of what type of drugs will be helpful or best suited for an individual patient [37, 38].

Take the example of lung cancer, where earlier the treatment was performed on the basis of the type of tissue that patient has rather than focusing on the specific genetic mutation [16]. Now, taking gene sequencing into account, healthcare experts can now recognize the genetic mutations which are the root cause of the condition, and not just do the treatment based on the symptoms but can provide a holistic treatment.

As every cancer is caused by mutations in genes [15, 16]. Hence, if we identify which genetic mutation will be involved in the growth of the tumor, then it would be possible to detect cancer at an early stage and it will provide the direction for the best drug treatment to the physician. There is a huge amount of data related to gene mutation in the form of literature text, gene expression form and many more.

The goal is to understand the genetic data to detect cancer. As the complexity and ever increasing amount of unstructured data, artificial intelligence (AI) can extensively be applied to achieve this goal.

2.2 Proposed Methodology

2.2.1 Introduction

The sequence of nucleotides forms a gene in RNA/DNA. The gene product (RNA or protein) is encoded by DNA or RNA. Sugar, Phosphate and Nitrogen Base are the components that form the nucleotide. The five bases of DNA/RNA are A, G, C, T and U. Genetic variations in humans are generally called the difference in DNA [20].

Mutation and genetic recombination are two of the most common sources of genetic variation. The change in nucleotide sequence of the genome is called the mutation. Cancer post sequencing can have hundreds of genetic mutations [17, 18]. The task of identifying the genetic mutation is challenging and time consuming since this is done manually most of the time. The factors such as heterogeneity, stage of cancer, treatment options and patient country pose a challenge in precise prediction of survival in cancer patients.

Currently, the decision on which treatment to use is taken from the data collected at medical records, cancer registries and bedside consultations, gene expression data. If these predictions can be achieved more reliably, the doctors can intervene in a much better way and improve the overall performance of institutions in cancer management.

Therefore, there is a need to understand the collected data efficiently and also need to detect the cancer at an early stage. As Cancer doesn't happen due to a single mutation, it happens due to a large number of mutations over a human lifetime [22]. Older people hence are more likely to get cancer as they have more time for mutations to build up.

With the advanced capabilities of AI, healthcare professionals can have the facility to break down the complex problems which have complicated patterns to identify, time taking or hard to tackle alone. AI can result to be a valuable resource in the domain of healthcare which will allow health professionals to use their experience and knowledge to its best level. Considering the same, this research focused on to propose a methodology which will help to understand the data and identify the complex pattern from the data so that cancer classification can be done accurately.

This study has been done to classify cancer based on gene expression data. Gene expression data is the representation of microarray images in matrix format. Microarray has the capability to simultaneously monitor ten thousand expressions of genes which amounts to informative data. In gene expression data, genes are represented in rows and columns are represented by varied samples of tissues or external factors. The number in a particular cell characterizes the expression level of the distinct gene in the distinct samples. To extract the useful and necessary information from this data and use it for classification of cancer, we have proposed the methodology which has been shown in figure 2.1.

The first step is data preprocessing which has been used to normalize the features of datasets. Further, feature transformation has taken place which is done by autoencoder technique. Autoencoder will get trained on the given dataset. Once it will get trained, the encoder part will be stacked to the deep neural network to classify the cancer class.

2.2.2 Dataset Description

Cancer is known to be the one of the main reasons for deaths worldwide. There are more than hundred types of cancer and all are the results of unwanted mutation happening in cells. Researchers work to find the difference between normal and

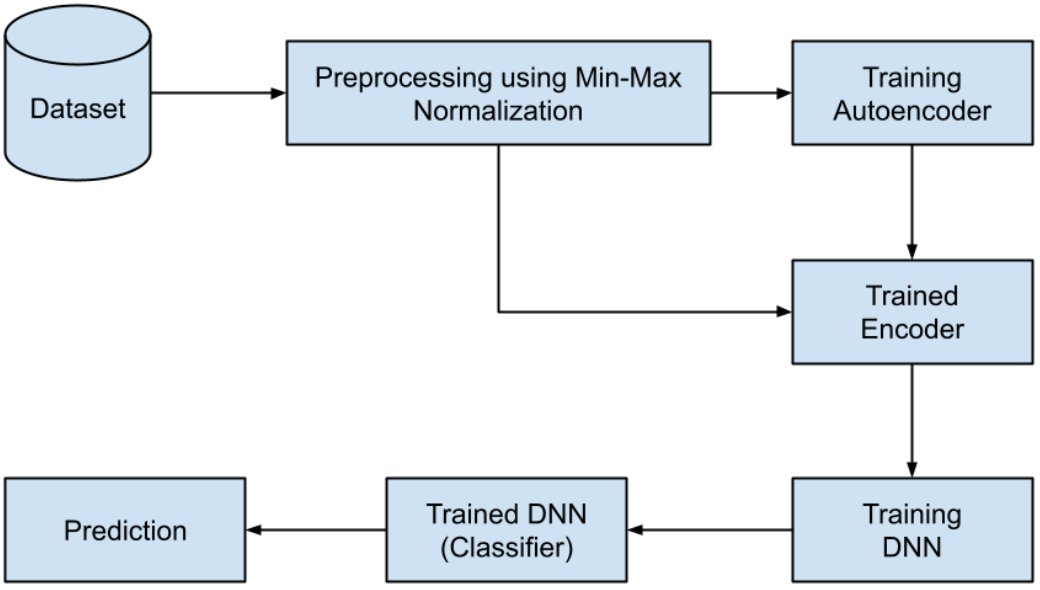


FIGURE 2.1: Architecture of Proposed Methodology.

unhealthy cells. The major concern is to detect those genes which would be involved in the growth of tumors. Hence, there are a large number of datasets available which have information like literature text related to genetic mutation, gene expression data, epigenetic and genetic cytology. This work has used gene expression data to classify cancer. There are several datasets available based on gene expression which includes leukemia, DLBCL, BC-TCGA, prostate cancer, breast, lung, colon tumor and NC160. In this research, the work is done on four gene expression datasets whose description has been discussed below. All the datasets are freely available and got from Kent Ridge Bio-medical Dataset [39].

- Leukemia Dataset : This dataset has been provided by Golub et. al.(1999). The sample has been collected from leukemia patients where some of them are ALL (Acute Lymphoblast Leukemia) and some of them are AML (Acute Myeloid Leukemia). It contains gene expression which has been taken from bone marrow and peripheral blood samples. Each expression level has been measured by 7,129 genes.

- Postate Dataset : This dataset has been provided by Singh et al. The sample has been collected from prostate cancerous patients and normal healthy persons. Each sample has been measured by 12533 gene expression levels.

- DLBCL Dataset : This dataset has been used in the research done by Shipp et al. The collected data has been taken from DLBCL (Diffuse Large B-cell Lymphoma) and from Follicular lymphoma (FL). Each sample has been measured by 5469 expression levels.

- Colon tumor Dataset : The dataset has been used in the research done by Alon et al. The sample has been taken from cancerous biopsies and healthy biopsies. Each sample has been measured by 2000 gene expression levels.

2.2.3 Data Preprocessing

Data preprocessing is one of the key steps in training a neural network or machine learning model as the quality of the data often determines the ability of the model to learn. It is a significant step in the entire orchestration process. The performance of the model has been based on the quality of data provided for learning. So, it is important to refine the data prior to feeding it to the model. In this thesis, data normalization has been done on all datasets.

Example : Let's say in houses dataset total number of rooms and age of house are two important features. We are trying to predict which house would be the best for us. The feature on a bigger scale dominates the other feature when the model is trying to compare the data points.

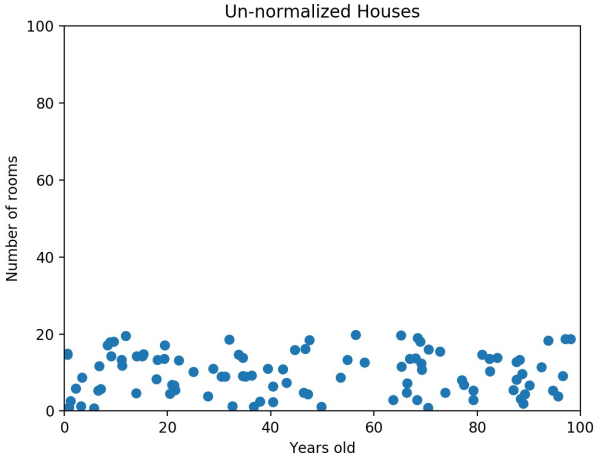


FIGURE 2.2: Unnormalized datapoints [5]

The problem we have now is that the data is squished as shown in figure 2.2. Machine learning models should be able to understand that there is a difference between a house with 20 rooms vs 2 rooms. However since these two houses are a hundred years apart the contribution made by difference in rooms is very less. This could be solved by normalization.

2.2.3.1 Min-Max Normalization

Normalization is done to scale the features at the same level. One of the normalization techniques is min-max normalization where the feature with minimum value is replaced by 0 and feature with maximum value replaced with 1. The remaining features are replaced by the decimal value which will be in the range of 0 and 1. The following formula is used to evaluate the normalized value for features.

$$Norm(k) = (k - min\_val) / (max\_val - min\_val) \tag{2.1}$$

where K denotes the feature value, max_value tells the maximum value among the features and min_value represents the minimum value among the features.

Input for Min Max Normalization : Let's say we are normalizing the number of rooms first and let's say we have (3, 1, 5, 2) in the dataset. Here the minimum value is 1 and maximum value is 5. We now iterate over the dataset and divide the difference between the current datapoint and minimum value with the difference between minimum value and maximum value. This kind of normalization is called min-max normalization and it is one of the most popular ways to normalize the data.

Output for Min Max Normalization : A single scalar value between 0 and 1.

The whole aim of this normalization process is to make every data point be on the same scale so every feature will be equally important as shown in figure 2.3.

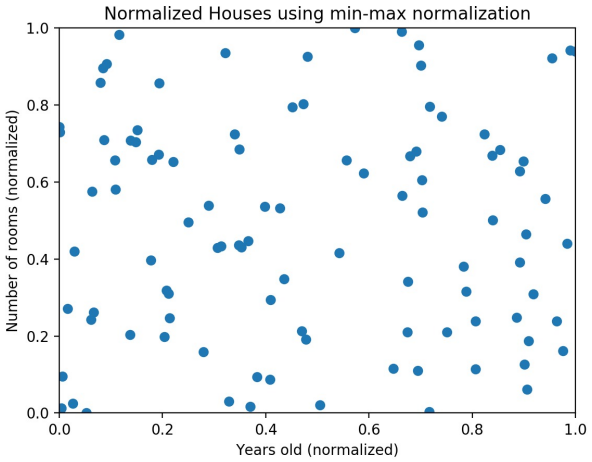


FIGURE 2.3: Normalized datapoints [5]

2.2.4 Feature Extraction from raw data

2.2.4.1 Autoencoder

In the modern computing era, Deep learning and Artificial Intelligence have made huge progress in NLP and computer vision. Due to this reason deep neural networks have been limited to image data or sequential data(documents). So data in tables/csv/dataframes is limited to only simple multi level perceptron models. RNN's are used when the problem involves tabular data with time series, however the usage of RNN is also limited to only time series data. Without time series dataset type there isn't any other use case where a deep neural network is used on tabular data. So for the purpose of this thesis we've used Autoencoder on non time-series tabular data, however autoencoders are generally used in image or text data.

Autoencoders as a feature transormation technique : If we have a dataset where the number features are very high then we can trim the total number of features by using autoencoders through non-linear complex functions that are abstracted out to practitioners. Using autoencoders is a better way for dimensionality reduction compared to other traditional methods like principal component analysis [40].

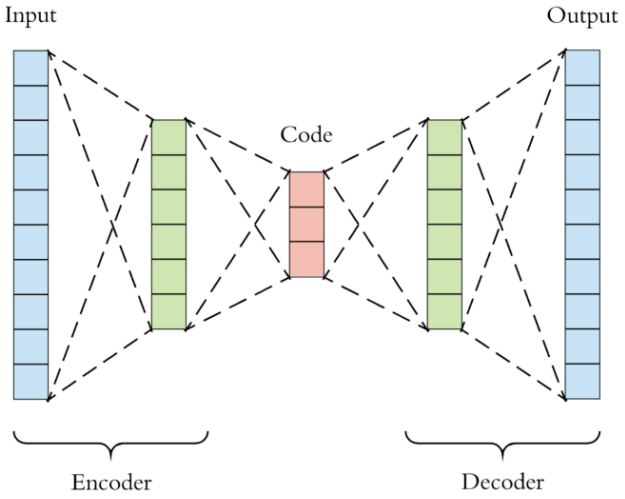


FIGURE 2.4: Structure of AutoEncoder [6]

Autoencoders predict the output directly without looking at the labels or feature names in the dataset. The autoencoder generally has three parts as shown in figure 2.4.

- Encoder
- Bottleneck
- Decoder

We can use MSE or MAE when training the autoencoder. If the given data is X and the output given by the autoencoder is x_out, we try to optimize.

$$L(X, X\_out) = |X - X\_out|^2 \tag{2.2}$$

Once the autoencoder is trained, we can use the encoder part of the autoencoder for dimensionality reduction.

2.2.5 Stacked Autoencoder-DNN Model

Deep neural networks is the most promising technique of the current generation of Artificial Intelligence. We have taken the DNN architecture proposed in the baseline model and stacked an encoder before it so that the encoder filters out the

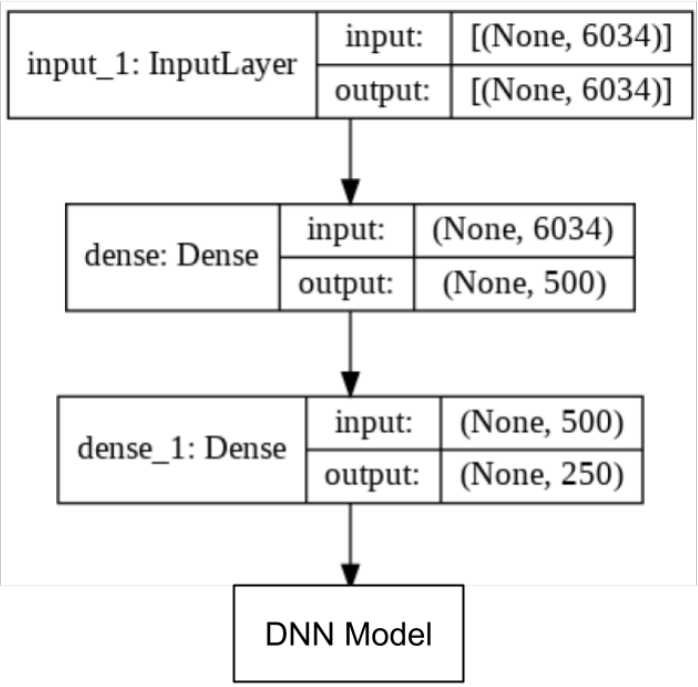


FIGURE 2.5: Architecture of Stacked Autoencoder-DNN Model

noise [3]. The architecture which is shown in figure 2.5 is giving empirically better results compared to the original model proposed in the base paper.

The model which is shown in figure 2.6 consists of an auto encoder which is trained on the gene expression dataset so that it learns how to encode and decode. Once the model has successfully trained on how to encode data into lower dimensions and decode from lower dimensions. We will use the encoder part of the autoencoder to encode the data and then give it to a deep neural network.

The deep neural network takes the output of the encoder and trains itself to perform classification tasks. The architecture of DNN is shown in figure 2.7 which consists of 256 neurons at the input layer followed by hidden layers with 64, 32, 16 and 1 neurons. There is a dropout of 20% of the nodes to perform regularization. Autoencoder will have N number of neurons at the input layer if there are N features followed by X number of neurons at the bottleneck layer if we want to reduce the dimensions to X. We have performed a lot of hyperparameter tuning to come up with the optimal number of dimensions at the bottleneck layer.

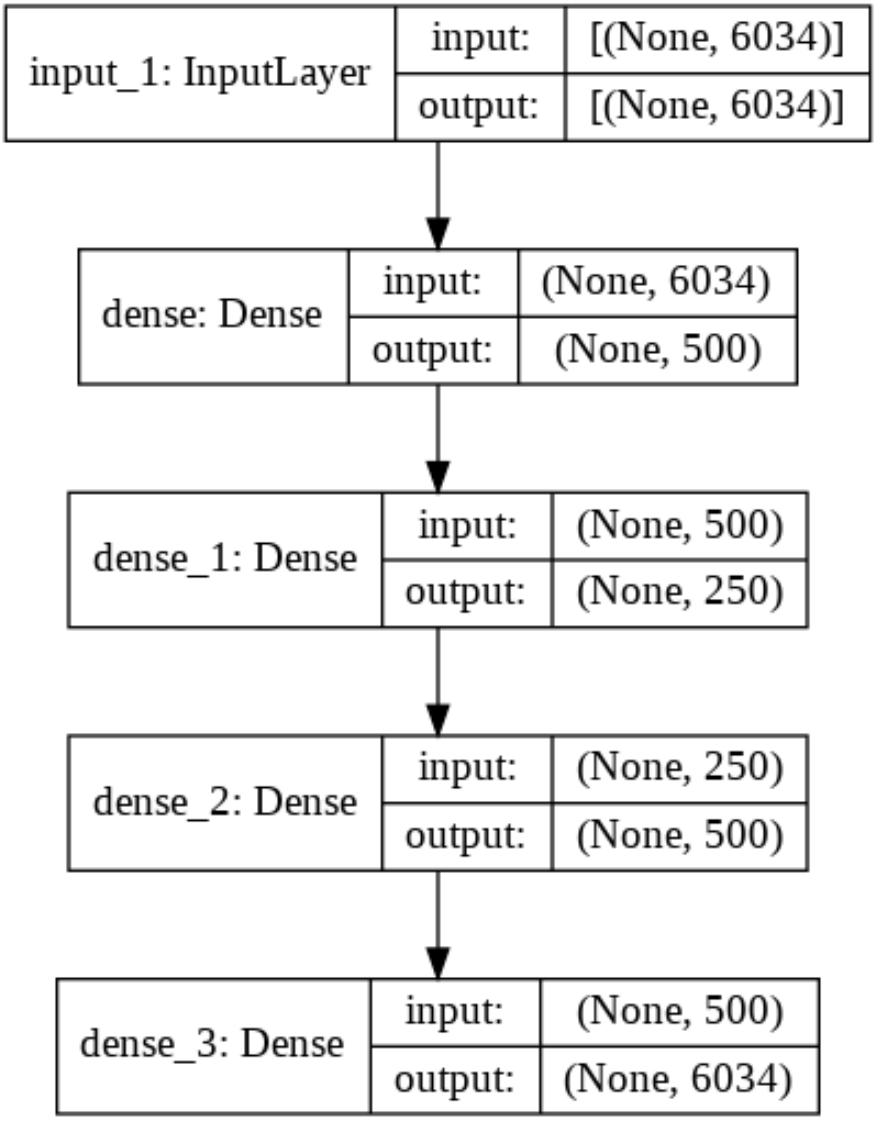


FIGURE 2.6: Architecture of Autoencoder

2.3 Conclusion

In this chapter, we gave an overview of the proposed approach for cancer classification based on gene expression data.

Here, we explained the Min Max normalization and why there is a need to include this in our study. We have discussed model architecture and elaborated its components which includes Auto encoder and DNN model.

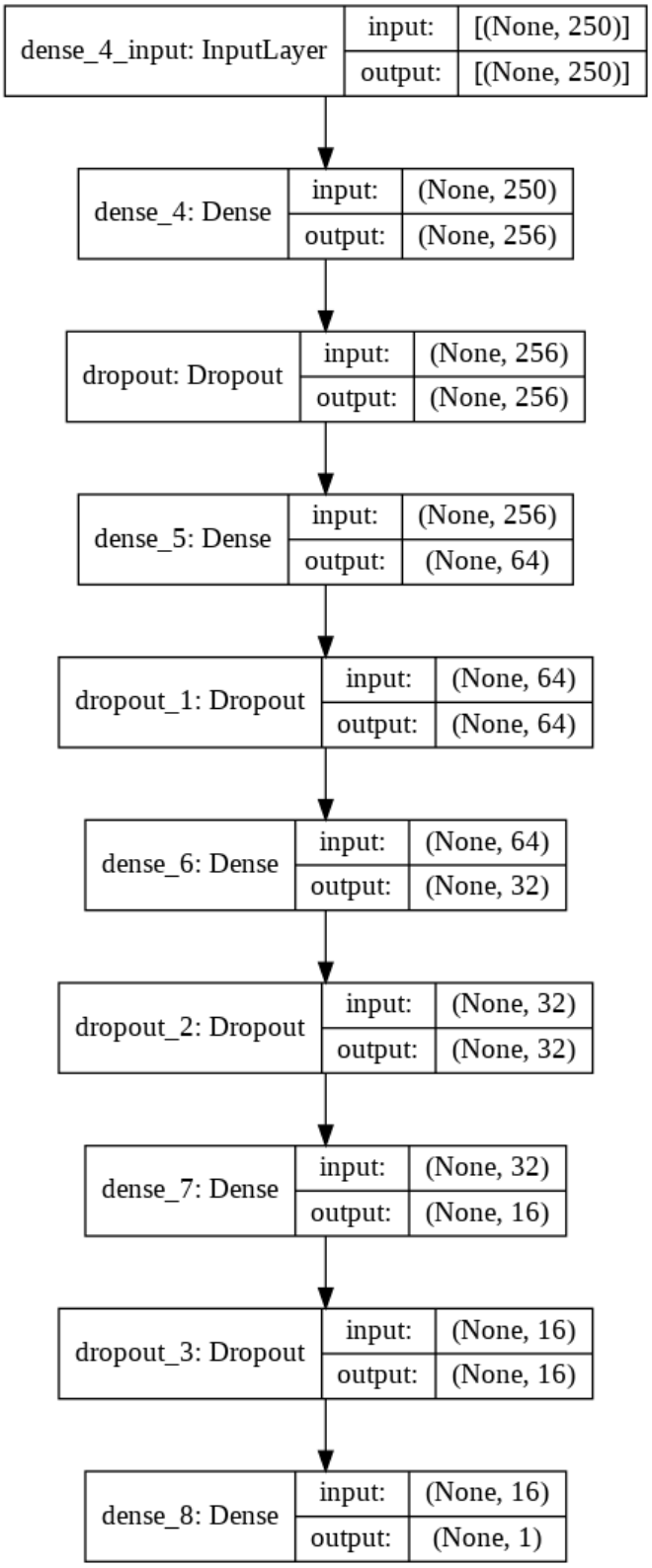


FIGURE 2.7: Architecture of Deep Neural Network

In the next chapter, we will take a look at the results obtained by the proposed approach, the libraries and platforms used to implement the approach and at the end we will perform an extensive comparison analysis.

Chapter 3

Experimental results and comparison

3.1 Implementation of Proposed Work

3.1.1 Software Used

3.1.1.1 Python 3.7

In this proposed work, Python 3.7 has been used. Python 3.7 came up with many new libraries irrespective of it's previous versions [41]. It is quite easy to use these libraries according to our requirements. That's why it is a highly acceptable language for machine learning work. The installation and setup of the platform is quite easy which provides ease to the programmer to make it use for the work. The platform has modules and their implementation is quite easy that reduces the lots of work of programmers.

3.1.2 Libraries Used

3.1.2.1 Pandas

Panda is highly compatible for data analysis. It acts as a wrapper and contains the function that helps the programmer to find the meaningful, non redundant and logical data. Functions such as groupby() are used to split and combine the data logically. There are two data structures named Series and Data Frames. We have used Dataframes in this work as it is used to store the data in tabular format and also helps to read the data from csv files. This is easy to grasp which makes it used widely [42].

3.1.2.2 Numpy

As the name suggests Numeric Python, it is helpful in computation of arrays and matrices. It offers many computational functions that make scientific calculations on vectors very easy [43]. It provides high end tools to work with arrays. So, we used this to compute the array in our work to solve typical mathematical problems and this makes it easy to analyze the numerical data.

3.1.2.3 Sklearn

Sklearn stands for scikit-learn is the python library. The tools are very helpful for machine learning as it creates the module of the work [44]. The library has many algorithms which help to perform different tasks such as to classify and cluster the data. It also regresses the data that create the relevant module.

3.1.2.4 Keras

Keras is a high level API that offers the easiest ways to implement a deep neural network [45]. It is designed to perform fast and different experiments on neural networks without compromising the focus on being convenient, compatible and expandable. It supports various neural networks including RNN, CNN and the

combination of both. Keras offers tools that can work with pictorial and textual data easily hence making it simple for us to write neural network code. It can run on both CPU and GPU and thus extensively used in the field of Deep Learning. It acts as a simple interface for the Tensorflow library making it a user-friendly way to create networks.

3.1.3 Platform Used

3.1.3.1 Intel Devcloud JupyterLab

Intel Devloud is a scalable processor that helps to compute the data of machine learning and deep learning.This makes it easy to stage and compile the code. Jupyter Notebook combines with DevCloud to create a platform where developers can write code in Python, Julia and many other languages. In our work we have written the code instructions in Python. Intel Devcloud Jupyter Lab gives the feeling of a modern Integrated Development Environment where developers can build the application with a common graphical user interface [46].

3.2 Results and Analysis

For the purpose of this thesis, cancer classification has been done on four different datasets which are based on gene expression level. The proposed model is an autoencoder based deep neural network with dropout and shows improved results than the existing machine learning models. In the existing work, the author used several algorithms which include DNN, CNN, RNN and improved DNN (dropout on the base DNN model) [3]. Improved DNN / DNN with dropout has shown the best results among all the other algorithms on these four datasets. However, the results do not provide good results on Prostate and Colon tumors dataset as compared to Leukemia and DLBCL dataset.

In the proposed approach, we have used a stacked Autoencoder-DNN with dropouts model and preprocessed the data by normalizing the features using min-max normalization. The details of training DNN model has been given in table 3.1.

There are some features of gene expression data which includes the high dimensionality, smaller sample size, and not all the genes present in data are relevant to that tumor. So most of the genes are irrelevant and some of them are useful. It is required to get relevant data. Autoencoder is a feature transformation technique which helps to extract relevant features and feeding relevant features to model, helps to learn model more efficiently and hence we have seen slight improvement in accuracy by using classical autoencoder.

TABLE 3.1: Details of training the DNN model

Attributes	**Value**
Activation Function of all layer	Relu function and sigmoid function at output layer
Type of internal layer unit	Simple Dense Unit
Loss Function	Binary Cross Entropy Loss
Optimizer (learning rate)	Adam
Initial Weights	Random

TABLE 3.2: Comparison of Classification Accuracy of Proposed approach with Existing approach

Dataset	**DNN with dropout [3]**	**Stacked Autoencoder-DNN Model**
Prostate	93.2	95.2
DLBCL	98.4	99.3
Leukemia	99	99.1
Colon Tumor	91.4	94.8

As shown in table 3.2, the proposed approach has provided 95.2% accuracy on Prostate dataset followed by 99.4% accuracy on Leukemia, 94.8% on Colon Tumor dataset and 99.2% on DLBCL dataset. The results are improved on all the datasets as compared to the existing base model.

The performance metric used in this work is 'Accuracy' which is the ratio of the total number of points which are correctly classified to the total number of points in the testing data.

Chapter 4

Conclusion

4.1 Conclusion and Future Work

This research has focussed on cancer classification based on gene expression data using deep learning algorithms. Although deep learning is very well established for over several decades in the healthcare domain, the DL approach has been used lately to categorise whether a patient will acquire a particular disease or not. It formulates problems on the basis of input terms, output terms and the other dependent variables that link inputs with outputs. It is extensively useful to identify complex patterns from raw data.

The proposed methodology has used a DNN stacked with an autoencoder model that has been trained and tested on four different datasets including Prostate, Leukemia, DLBCL and Colon tumor. The results showed better performance on all the dataset in comparison to DNN with dropouts only.

Every year new types of mutations occur and lack of those mutations in the training dataset will lead to poorer results for that particular mutation, so future work would include continuously upgrading the data with all the new mutations worldwide and incorporating those mutations into newer deep learning approaches that emerge with time.

Bibliography

[1] Researchers have discovered a mathematical relationship that sheds new light on the rate at which cancer cells mutate - https://debuglies.com. https://bit.ly/cancer-cells-mutate.

[2] Support vector machines (svm) — learn opencv. https://learnopencv.com/support-vector-machines-svm/.

[3] Omar Ahmed and Adnan Brifcani. Gene expression classification based on deep learning. In *2019 4th Scientific International Conference Najaf (SICN)*, pages 145–149. IEEE, 2019. doi: 10.1109/SICN47020.2019.9019357. URL https://ieeexplore.ieee.org/document/9019357.

[4] Jain K. Goyal P., Pandey S. In *Unfolding Recurrent Neural Networks*. In: Deep Learning for Natural Language Processing. Apress, Berkeley, CA, 2018. URL https://doi.org/10.1007/978-1-4842-3685-7_3.

[5] Min max normalization. https://www.codecademy.com/articles/normalization.

[6] Unconventional deep learning techniques for tabular data — deep learning for tabular data. https://bit.ly/dont-stop-at-ensembles-unconventional-deep-learning-techniques.

[7] Douglas Hanahan. Rethinking the war on cancer. *The Lancet*, 383(9916):558–563, 2014. ISSN 0140-6736. doi: https://doi.org/10.1016/S0140-6736(13)62226-6. URL https://www.sciencedirect.com/science/article/pii/S0140673613622266.

[8] Robert A. Weinberg. Special issue: What you need to know about cancer. In *How Cancer Arises.*, volume 275, pages 62–70. Scientific American, a division of Nature America, Inc., 1996. URL http://www.jstor.org/stable/24993349.

[9] Xiangming Guan. Cancer metastases: challenges and opportunities. *Acta Pharmaceutica Sinica B*, 5(5):402–418, 2015. ISSN 2211-3835. doi: https://doi.org/10.1016/j.apsb.2015.07.005. URL https://www.sciencedirect.com/science/article/pii/S2211383515001094.

[10] Helen Dimaras, Kahaki Kimani, Elizabeth AO Dimba, Peggy Gronsdahl, Abby White, Helen SL Chan, and Brenda L Gallie. Retinoblastoma. *The Lancet*, 379(9824):1436–1446, 2012. ISSN 0140-6736. doi: https://doi.org/10.1016/S0140-6736(11)61137-9.

[11] cancer research uk 2021. Types of cancer. https://www.cancerresearchuk.org/what-is-cancer/how-cancer-starts/types-of-cancer, May 2021.

[12] Erik Sahai Peter Friedl, Joseph Locker and Jeffrey E. Segall. Classifying collective cancer cell invasion. *Nature Cell Biology*, 14:777–783, 2012. doi: 10.1038/ncb2548. URL https://doi.org/10.1038/ncb2548.

[13] Stewart Sell. Leukemia. *Stem Cell Reviews*, 1:197–205, 2005. ISSN 1558-6804. doi: 10.1385/SCR:1:3:197. URL https://doi.org/10.1385/SCR:1:3:197.

[14] Maria Bach Laursen, Steffen Falgreen, Julie Støve Bødker, Alexander Schmitz, Malene Krag Kjeldsen, Suzette Sørensen, Jakob Madsen, Tarec Christoffer El-Galaly, Martin Bøgsted, Karen Dybkær, and Hans Erik Johnsen. Human b-cell cancer cell lines as a preclinical model for studies of drug effect in diffuse large b-cell lymphoma and multiple myeloma. *Experimental Hematology*, 42(11):927–938, 2014. ISSN 0301-472X. doi: 10.1016/j.exphem.2014.07.263. URL https://doi.org/10.1016/j.exphem.2014.07.263.

[15] Peter A. Jones and Stephen B. Baylin. The epigenomics of cancer. *Cell*, 128 (4):683–692, 2007. ISSN 0092-8674. doi: 10.1016/j.cell.2007.01.029. URL https://doi.org/10.1016/j.cell.2007.01.029.

[16] The genetics of cancer. https://www.cancer.net/navigating-cancer-care/cancer-basics/genetics/genetics-cancer, Jun 2021.

[17] The genetics of cancer - national cancer institute. https://www.cancer.gov/about-cancer/causes-prevention/genetics, April 2015.

[18] Suzuki DT et al Griffiths AJF, Miller JH. Mutation and cancer. *An Introduction to Genetic Analysis*, 7th edition, 2000. URL https://www.ncbi.nlm.nih.gov/books/NBK21809/.

[19] Yarbro JW. Oncogenes and cancer suppressor gene. *An Introduction to Genetic Analysis*. doi: 10.1016/0749-2081(92)90006-o.

[20] R. A. Sclafani and T. M. Holzen. Cell cycle regulation of dna replication. *Annual Review of Genetics*, 41(1):237–280, 2007. doi: 10.1146/annurev.genet.41.110306.130308. URL https://doi.org/10.1146/annurev.genet.41.110306.130308. PMID: 17630848.

[21] Holtedahl K. Challenges in early diagnosis of cancer: the fast track. pages 251–252, 2020. doi: 10.1080/02813432.2020.1794415. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7470137/.

[22] Tania Estapé. Cancer in the Elderly: Challenges and Barriers. pages 40–42, 2018. doi: 10.4103/apjon.apjon_52_17. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5763438/.

[23] Cancer Research UK. Why Is Early Diagnosis Important? 2015. URL https://www.cancerresearchuk.org/https%3A//www.cancerresearchuk.org/about-cancer/spot-cancer-early/why-is-early-diagnosis-important.

[24] Alvis Brazma and Jaak Vilo. Gene expression data analysis. *FEBS Letters*, 480(1):17–24, 2000. ISSN 0014-5793. doi: https://doi.org/10.1016/S0014-5793(00)01772-5. URL https://www.sciencedirect.com/science/article/pii/S0014579300017725. Functional Genomics.

[25] Parmigiani Giovanni et al. The analysis of gene expression data: An overview of methods and software. *Springer New York*, page 1–45, 2003.

doi: 10.1007/0-387-21679-0_1. URL https://www.sciencedirect.com/science/article/pii/S0014579300017725.

[26] Downes C.S. Berrar D. Dubitzky W., Granzow M. Introduction to microarray data analysis. *A Practical Approach to Microarray Data Analysis*. doi: 10.1007/0-306-47815-3_1. URL https://doi.org/10.1007/0-306-47815-3_1.

[27] Adi L. Tarca, Roberto Romero, and Sorin Draghici. Analysis of microarray experiments of gene expression profiling. *American Journal of Obstetrics and Gynecology*, 195(2):373–388, 2006. ISSN 0002-9378. doi: 10.1016/j.ajog.2006.07.001. URL https://doi.org/10.1016/j.ajog.2006.07.001.

[28] Joseph M. De Guia. Cancer classification of gene expression data using machine learning models. *2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology,Communication and Control, Environment and Management (HNICEM)*, pages 1–6, 2018. doi: 10.1109/HNICEM.2018.8666435. URL https://ieeexplore.ieee.org/document/8666435.

[29] Principal component analysis for dimensionality reduction — by lorraine li — towards data science. https://towardsdatascience.com/principal-component-analysis-for-dimensionality-reduction-115a3d157bad.

[30] Suthaharan S. Support vector machine. In *Machine Learning Models and Algorithms for Big Data Classification*, volume 36. Integrated Series in Information Systems, 2016. doi: 10.1007/978-1-4899-7641-3_9. URL https://doi.org/10.1007/978-1-4899-7641-3_9.

[31] Xgboost - wikipedia. https://en.wikipedia.org/wiki/XGBoost.

[32] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. ACM International Conference on Knowledge Discovery and Data Mining, (SIGKDD'16), San Francisco, California, USA, 2016.

[33] Nicolas Lachiche Peter A Flach. In *Naive Bayesian Classification of Structured Data*, pages 233–269. Machine Learning, 2004. doi: 10.1023/B:MACH.0000039778.69032.ab. URL https://doi.org/10.1023/B:MACH.0000039778.69032.ab.

[34] G. Liu H. Zeng, M. D. Edwards and D. K. Gifford. Support vector machine. In *Convolutional neural network architectures for predicting DNA–protein binding*, volume 32, pages 121–127. Bioinformatics, 2016. doi: 10.1093/bioinformatics/btw255.

[35] A. Mohamed A. Graves and G. Hinton. In *Speech recognition with deep recurrent neural networks*, number 12, page 6645–6649. IEEE International Conference on Acoustics, Speech and Signal Processing, 2013.

[36] Weibo Liu, Zidong Wang, Xiaohui Liu, Nianyin Zeng, Yurong Liu, and Fuad E. Alsaadi. A survey of deep neural network architectures and their applications. *Neurocomputing*, 234:11–26, 2017. ISSN 0925-2312. doi: 10.1016/j.neucom.2016.12.038. URL https://doi.org/10.1016/j.neucom.2016.12.038.

[37] Sophia genetics - wikipedia. https://en.wikipedia.org/wiki/Sophia_Genetics.

[38] Sophia genetics - where others see data we see answers - where others see data we see answers. https://www.sophiagenetics.com/.

[39] In *Leukemia and Colon Tumor dataset*. URL https://leo.ugr.es/elvira/DBCRepository/.

[40] Wei Wang, Yan Huang, Yizhou Wang, and Liang Wang. Generalized autoencoder: A neural network framework for dimensionality reduction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2014.

[41] Python 3.7 documentation — devdocs. https://devdocs.io/python~3.7/. (Accessed on 07/11/2021).

[42] Python pandas - introduction - tutorialspoint. https://www.tutorialspoint.com/python_pandas/python_pandas_introduction.htm.

[43] Introduction to numpy. https://www.w3schools.com/python/numpy/numpy_intro.asp.

[44] An introduction to machine learning with scikit-learn — scikit-learn 0.24.2 documentation. https://scikit-learn.org/stable/tutorial/basic/tutorial.html.

[45] Introduction to keras. https://pypi.org/project/keras/.

[46] Base training modules — intel® devcloud. https://devcloud.intel.com/oneapi/get_started/baseTrainingModules/.

[47] How cancer can spread — cancer research uk. https://www.cancerresearchuk.org/about-cancer/what-is-cancer/how-cancer-can-spread.