EAST WEST UNIVERSITY

Somatic Mutation Prediction in Human DNA Sequence in Absence of Matching Samples Using Artificial Neural Network

Submitted By

Pritam Datta

ID: 2014-2-60-078

Md. Sakib Hasan

ID: 2014-2-60-077

Md. Abdur Rahaman

ID: 2014-2-60-082

Supervised By

Dr. Maheen Islam

Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

EAST WEST UNIVERSITY

This project has been submitted to the Department of the Computer Science & Engineering at East West University in the partial fulfillment of the requirement for the degree of Bachelor of Science in Computer Science and Engineering.

August , 2018

Abstract

Somatic mutation can occur at any stages of an individual which may turn towards cancer due to normal genome cell transformations. Hence, its early detection is essential for the appropriate treatment and other purposes. Generally, mutated genes are matched with the normal tissues of the donor. In addition, mutated genes are compared with the existing publicly available mutational loads. However; this computation is too costly due to appropriate matching samples and in terms of computational complexity, time consumption, memory usage etc. In these consequences, this paper proposes an efficient machine learning based approach to distinguish the somatic single nucleotide variants in absence of matching samples. Here, we have applied multilayer perceptron of artificial neural network on the standard training sets like BRCA, COAD, PAAD, KIRC, ESC, UCEC. Then the results are validated using 10-fold cross validation technique. The maximum accuracy active by the execution of the proposed scheme is 97% with f1-measure ranges from 88-97% for different cancer types which is much higher than the existing state of the art approaches.

Declaration

We hereby, declare that all the work presented in this project is the outcome of the investigation and research performed by us under the supervision of Dr. Maheen Islam, Assistant Professor, Department of Computer Science and Engineering, East West University, Dhaka, Bangladesh. We also declare that neither it nor part of it has been submitted for the requirement of any degree or diploma or for any other purposes except for publications.

Signature of the Candidate

… … … … … … …	**… … … … … … …**	**… … … … … … …**
Pritam Datta	**Md. Sakib Hasan**	**Abdur Rahman**
ID: 2014-2-60-078	**ID: 2014-2-60-077**	**ID: 2014-2-60-082**

Letter for Acceptance

This Project entitled "Somatic Mutation Prediction in Human DNA Sequence in Absence of Matching Samples Using Artificial Neural Network" submitted by Pritam Datta (ID:2014-2-60-078), Md. Sakib Hasan (ID:2014-2-60-077) and Md. Abdur Rahaman (ID:2014-2-60-082) to the Department of Computer Science and Engineering, East West University, Dhaka, Bangladesh is accepted by the department in partial fulfillment of requirements for the Award of the Degree of Bachelor of Science in Computer Science and Engineering on August, 2018.

Supervisor

____

Dr. Maheen Islam
Assistant Professor
Department of Computer Science and Engineering,
East West University, Dhaka, Bangladesh

Chairperson

____

Dr. Ahmed Wasif Reza

Associate Professor and Chairperson,

Department of Computer Science and Engineering,

East West University, Dhaka, Bangladesh

Acknowledgements

First, we are thankful and expressing our gratefulness to Almighty who offered us divine blessings, patience, mental and psychical strength to complete this project. The progression of this thesis could not possibly be carried out without the help of several people who, directly or indirectly, are responsible for the completion of this work. We deeply indebted to our project supervisor Dr. Maheen Islam. Her scholarly guidance, especially for her tolerance with our persistent bothers and unfailing support. She gives us the freedom to pursue aspects of Artificial neural network in somatic mutation which we found interesting and compelling. This helped our thesis to achieve its desired goals.

We wish to thank the great people of Department CSE at East West University. A special thank goes to all faculties for their well-disposed instructions and Encouragements.

Finally, we would like to thank our friends and family. Their continued tolerance with our moods and tendency to disappear for weeks at a time gave us a much needed break from the world computing.

Table of Contents

Chapter 1

Introduction

Somatic mutation.is alternation in the genetic structure which is not inherited from a parent, and also not passed to offspring, is called a somatic mutation. In any stages of a human this mutation can occur and resulted in cancer.
Identification of Somatic mutation is the vital key step for many cancer related studies [2]. Among many challenges in mutation calling, some of the challenges include:

Admixture of multiple tumor sub-clones (among them and with other normal tissue), Multiple number of presences of copy number alterations in tumor, Raw error rate that generates from sequencing instruments and also which is comparable to the mutant allele's variant allele frequency in admixed samples etc.

Many of the mathematical models and computational approaches focus on the mutational pattern as well as cancer genome. There are many computational approaches designed to identify somatic mutation prediction such as the help of gene expression integrating somatic mutation calls, Using known pathways (biological pathways) from public database etc. Among them one of the computational approaches is the use of cross-validation. This is done for unraveling the somatic events. Here, in this case the goal is to distinguish whether it is somatic or germline mutation in the absence of patient's matched normal samples. But if matched normal control is not required than there is an approach called massively parallel sequencing (MPS). This doesn't required patient's matched normal to distinguish whether it is somatic or germline mutation.

1.1 Motivation

Current generation of somatic mutation calling tools are reasonably good but low in terms of f1-measure and dependency on external database is higher for missing data. However, these tools [17][18] requires a larger number of reads to find some additional information like both patient's tumor and normal tissues such as MuSE, SNV Sniffer etc. In case of distinguishing somatic mutations from uncommon germline polymorphisms, all these tools require both normal tissues, for example white blood cells or adjacent normal tissue in the tumor resection specimen and patient's tumor. These tools construct multiple alignment using normal tissues and patient's tumor cell. Then, using statistical models of sequencing error rates and base quality scores, these tools write down the columns of the alignment to identify tumor specific alterations to reduce false positive. False positive is very important in medical terms of view. Suppose, a person get positive result for cancer but actually the result is not positive, that may impact him/her as he/she have to undergo treatment and it can also backfires in his/her body and cost money. Computational gene prediction is essential for genome sequencing project. So, as it is essential if false positive rate is high then there will be a negative impact on actual final result.

Generally, normal samples were not collected or patient permission was obtained in a way that prevents from the examination of normal tissue or germline variants. This is most commonly seen when dealing with the past events or performing analysis on previous studies with human material from legacy bio banks, clinical trials and pathology archives, a strategy which can be required when making a rare cancer type or subtype, or these can be seen when executing secondary studies on clinical trials. Another scenario is, there is no information on the donor's normal genomes in cancer cell line but cancer cell line has been used as an experimental model. Here, financial considerations can be a biggest fact. Data storage are increased due to the sequencing of tumor and normal genome and hence computational requirements and cost also increases.

Thus the identification of somatic mutation from tumor tissues is a challenging task for the researchers in especially in absence of normal tissue samples.

In this paper, we have proposed an efficient method to identify somatic mutation without matching normal samples, the role of neural network to increase its performance, to implement computational approaches focusing on the mutational pattern as well as cancer genome which will reduce data storage, cost, computational requirements and time consumption with better result.

1.2 Aims and Objectives

The objectives of this study are summarized below:

- To identify somatic mutation from human genome data without matching normal tissue sample.
- To get reasonably good accuracy with better f1-measure without the help of external database for preprocessing the dataset.
- To identify the factors of the features of the dataset that are responsible for somatic mutation and germline mutation.

1.3 Overview

This paper presents the identification of somatic mutation from tumor tissues in absence of normal tissue samples. We have proposed a supervised machine learning algorithm to distinguish between somatic mutation and germline mutation without the help of external database for preprocessing data and get good accuracy with better f1-measure than other scientific approaches. We also describes the factors of the features of the dataset that are responsible for somatic mutation and germline mutation by using decision tree.

1.4 Methodologies of Research

While working on this research, the following important steps are followed:
First, understanding of somatic and germline mutation, its cause and effects in Human DNA, finding the cause of cancer for that, the basics of mutation, its synthesis, the basics of Human genomic external databases, its synthesis, various existing approaches and methods to reduce somatic and germline mutation along with the role of neural network in this, etc.

- Designing the somatic mutation prediction pipeline using machine learning approach, studying existing mutation prediction approaches, then analyzing the designs, working procedures, advantages and shortcomings. Reducing the complexity.
- Inventing and finding the rules and methods for somatic mutation prediction without matching normal samples. Establishing the ideology and novelty of the proposed method through brief theoretical explanations. Comparing the findings based on the approach and analyzing the result.
- Finally, Comparing the methods and results of different cancer types with the existing state of the art approaches through artificial intelligence.

1.5 Outline

The next chapter (Chap.2) briefly discusses about the causes of mutation and the role of somatic and germline mutation in this. Also discusses about the effect of gene mutation in Human DNA.

Chap. 3 Discuss about the background study of identification and prediction of somatic and germline mutation.

Chap. 4 briefly introduces the genomic databases.

Chap. 5 discusses about the role and novel approaches of artificial neural network for somatic mutation prediction.

Chap. 6 Performance of proposed method is discussed here.

Chap. 7 finally discussed about Conclusions and Future work.

1.6 Summary

This chapter demonstrates motivations and objective of this thesis. Then the methodologies of the research that is being followed are discussed here. A brief elementary instructional text of remaining chapters of this thesis have also been described.

Chapter 2

Effect of GENE MUTATION in Human DNA

This chapter discuss about the basic definition and properties which are related with mutation and human DNA exome sequence. In this chapter, How somatic mutation occurs and how mutation can cause an effect on human DNA as well as the cause and effect of this will be discussed.

Gene mutation is a permanent alteration in the DNA sequence. Mutations can affect a single DNA building block (base pair) and a large segment of a chromosome that includes multiple genes. Generally there are two kinds' gene mutation (a) Somatic mutation and (b) Germline mutation [1].

2.1 Somatic mutation and its effect

The occurrence of a somatic mutation in the early stages of embryonic development or during the fusion of two zygotes. Somatic mutations are not inherited and do not affect the germline. These types of mutations are usually prompted by environmental causes. Somatic mutations are mutations that the genetic material of an organism acquires after its conception. They are called somatic due to their occurrence in the somatic (non-reproductive) cells of the organism. Since these mutations do not occur in the germ cells, they cannot be passed onto future progeny, and hence are also not inherited from the parent organism. Despite their non-hereditary nature, the risk of occurrence of a somatic mutation increases in the presence of other inheritable genetic factors and mutations. They are caused primarily due to environmental factors such as exposure to UV radiation, viral and bacterial infections, ingestion of toxic materials, faulty DNA repair, and unhealthy lifestyle choices[23].

A somatic mutation occurs after conception, after life starts. The occurrence of somatic mutations in a cell alter the genetic material of that cell and all the cells produced by its division, thus forming a mass of cells that possess a different genetic make up from that of the other cells of the organism. In some cases, if the mutations occur at a very early developmental stage, they lead to a condition where the organism may exhibit two or more sets of distinct sets of DNA throughout its body. Such an organism is called a "chimera". It is named after the Greek mythological creature of the same name, that was composed of parts taken from various different animals. A chimera is also formed when two fertilized eggs (zygotes) merge in utero to form a single embryo that has two sets of DNA. In some cases of chimerism, it has been observed that the mutated region of the body is later rejected by the rest of the body, and antibodies are produced against it.

Chimerism and somatic mutations are often mistaken for each other in case of alterations in phenotypes of an organism. In order to be sure of the cause, one must carry out DNA testing.

Spontaneously occurring mutations accumulate in somatic cells throughout a person’s lifetime. The majority of these mutations do not have a noticeable effect, but some can alter key cellular functions. Early somatic mutations can cause developmental disorders, whereas the progressive accumulation of mutations throughout life can lead to cancer and contribute to aging. Genome sequencing has revolutionized our understanding of somatic mutation in cancer, providing a detailed view of the mutational processes and genes that drive cancer.

Somatic mutations give rise to a variety of genetic disorders like Chimerism. The presence of a mutation in the somatic genes interrupts the sequencing and functions of the gene and the cells. The most common disease caused by somatic mutation is cancer, there are some other diseases caused by somatic mutation listed below[23]:

❑ Baraitser- winter syndrome
❑ Borching- Opitz syndrome
❑ Incontinentia pigment
❑ Kabuki syndrome
❑ Klippel- trenaunay syndrome
❑ Maffucci syndrome
❑ Neurofibromatosis
❑ Paroxysmal nocturnal hemoglobinuria
❑ Proteus syndrome
❑ Schinzel- Giedion syndrome

Example of somatic mutation:

A common example of somatic mutation can be seen in the dogs, variation of the color of the fur coat. When there is breeding in the dogs between same species then the puppy will have the color from one of the parent, but if there is somatic mutation then the there may be the inheritance of color of coat from both the parents.

For example, if a black German shepherd is mated with yellow German shepherd then the puppy may have the brown or yellow color but in the case of somatic mutation, there may be both colors present in the puppy.

2.2 Germline mutation and its effect

A germline mutation, or germinal mutation, is any detectable variation within germ cells (cells that, when fully developed, become sperm and ovum). Mutations in these cells are the only mutations that can be passed onto offspring, when either a mutated sperm or oocyte come together to form a zygote. Then fertilization event occurs, germ cells divide rapidly to produce all of the cells in the body, causing this mutation to be present in every somatic and germline cell in the offspring, this is also known as a constitutional mutation [3].

A germline mutation, occurs before conception, before life starts. It is present in the zygote. So every cell in the fetus will have it. The offspring of that fetus can inherit it. A gene change in a body's reproductive cell (sperm) that becomes incorporated into the DNA of every cell in the body of the offspring. Germline mutations are passed on from parents to offspring. Also called hereditary mutation. Different germline mutations can affect an individual differently depending on the rest of their genome. A dominant mutation only requires 1 mutated gene to produce the disease phenotype, while a recessive mutation requires both alleles to be mutated to produce the disease phenotype. For example, if the embryo inherits an already mutated allele from the father, and the same allele from the mother underwent an endogenous mutation, then the child will display the disease related to that mutated gene, even though only 1 parent carries the mutant allele. This is only one example of how a child can display a recessive disease while a mutant gene is only carried by one parent.

2.3 Rates of mutation

Mutations changes in the genetic sequence, and they are a main cause of diversity among organisms. These changes occur at many different levels, and they can have widely differing consequences. In biological systems that are capable of reproduction, we must first focus on whether they are heritable; specifically, some mutations affect only the individual that carries them, while others affect all of the carrier organism's offspring, and further descendants. For mutations to affect an organism's descendants, they must: 1) occur in cells that produce the next generation, and 2) affect the hereditary material. Ultimately, the interplay between inherited mutations and environmental pressures generates diversity among species.

Although various types of molecular changes exist, the word "mutation" typically refers to a change that affects the nucleic acids. In cellular organisms, these nucleic acids are the building blocks of DNA, and in viruses they are the building blocks of either DNA or RNA. One way to think of DNA and RNA is that they are substances that carry the long-term memory of the information required for an organism's reproduction. This article focuses on mutations in DNA, although we should keep in mind that RNA is subject to essentially the same mutation forces.

If mutations occur in non-germline cells, then these changes can be categorized as somatic mutations. The word somatic comes from the Greek word soma which means "body", and somatic mutations only affect the present organism's body. From an evolutionary perspective, somatic mutations are uninteresting, unless they occur systematically and change some fundamental property of an individual--such as the capacity for survival. For example, cancer is a potent somatic mutation that will affect a single organism's survival. As a different focus, evolutionary theory is mostly interested in DNA changes in the cells that produce the next generation[22].

Many direct and indirect methods have been developed to help estimate rates of different types of mutations in various organisms. The main difficulty in estimating rates of mutation involves the fact that DNA changes are extremely rare events and can only be detected on a background of identical DNA. Because biological systems are usually influenced by many factors, direct estimates of mutation rates are desirable. Direct estimates typically involve use of a known pedigree in which all descendants inherited a well-defined DNA sequence. To measure mutation rates using this method, one first needs to sequence many base pairs within this region of DNA from many individuals in the pedigree, counting all the observed mutations. These observations are then combined with the number of generations that connect these individuals to compute the overall mutation rate. Such direct estimates should not be confused with substitution rates estimated over phylogenetic time spans.

Mutation rates can vary within a genome and between genomes. Much more work is required before researchers can obtain more precise estimates of the frequencies of different mutations. The rise of high-throughput genomic sequencing methods nurtures the hope that we will be able to cultivate a more detailed and precise understanding of mutation rates. Because mutation is one of the fundamental forces of evolution, such work will continue to be of paramount importance[22].

2.4 Summary

This chapter shows the details description of somatic and germline mutation and how somatic mutation and germline mutation can causes effect in human genome. Also, how mutation can affect individual and other career organisms as well as how mutation rate can be measured is described here.

Chapter 3

Background study

This chapter discusses about the previous works of identification and the prediction of somatic mutation in human genome sequence as well as the previous works of artificial neural network in human DNA sequence.

3.1 SNP(Single nucleotide polymorphism)

SNP is single nucleotide polymorphism which is the most common genetic variation among people. It generally occurs in persons DNA.

Single-nucleotide polymorphisms may fall within coding sequences of genes, non-coding regions of genes, or in the intergenic regions (regions between genes). SNPs within a coding sequence do not necessarily change the amino acid sequence of the protein that is produced, due to degeneracy of the genetic code. SNPs in the coding region are of two types: synonymous and nonsynonymous SNPs. Synonymous SNPs do not affect the protein sequence, while nonsynonymous SNPs change the amino acid sequence of protein. The nonsynonymous SNPs are of two types: missense and nonsense. SNPs that are not in protein-coding regions may still affect gene splicing, transcription factor binding, messenger RNA degradation, or the sequence of noncoding RNA. Gene expression affected by this type of SNP is referred to as an eSNP (expression SNP) and may be upstream or downstream from the gene[24].

Inshort, SNP is a DNA sequence variation occurring when a single nucleotide adenine (A), thymine (T), cytosine (C), or guanine (G) in the genome (or other shared sequence) differs between members of a species or paired chromosomes in an individual. For example, two sequenced DNA fragments from different individuals, AAGCCTA to AAGCTTA.

On average, each human individual genome carried ∼3.3 million SNPs and ∼492,000 indels/block substitutions, including approximately 179 variants that were predicted to cause loss of function of the gene products [5]. Approximately numbers of coding variants are 20000 to 25000 in the genome of any human individual. And almost upto 9000 to 11000 are nonsynonymous which indicates nucleotide mutation that alters the amino acid sequence of a protein [6]. All SNPs which are common with 1% frequency have been largely catalogued in [7].. Some appropriate calibration within these groups may be needed and ethnic subpopulations are underrepresented. Moreover, every individual is estimated to carry rare SNPs of 400000 to 600000, specific to the individual or particular close family member as shown in [8]. But by comparison with SNP databases or by comparing with large scale exome sequencing projects, these cannot be removed easily. Here, one of the technical challenge is to have large number of storage capability. At least 4GB of RAM and 500GB of Hard Disk size is required to reformat and unzip all the external databases.

Next, Strelka , Safe mutation for deep and recurrent neural networks, and Identification of somatic mutation without matching normal tissue are reviewed.

3.2 Strelka somatic variant caller

To identify somatic or germline mutation with or without matching normal tissue such as there is a method called Strelka is which novel Bayesian approach is used for both tumor and normal samples to find somatic SNV and small indel from sequencing data. Strelka is an analysis package designed to detect somatic SNVs and small indels from the aligned sequencing reads of matched tumor-normal samples[19].

3.3 Safe mutation for deep and recurrent neural networks

Machine learning such as NN played a vital role for mutation in many methods such as a safe mutation for deep and recurrent neural networks which can help safe mutation operators that facilitate exploration without dramatically altering network behavior or requiring additional interaction with the environment[21]. The general approach for safe mutations is to choose weight perturbations in an informed way such that they produce limited change in the NN’s response to representative input signals.

The idea is to exploit sources of information that while generally are freely available, are often ignored and discarded. In particular, an archive of representative experiences and corresponding

NN responses can be gathered during an individual’s evaluation, which can serve to ground how dramatically a weight perturbation will changes the NN’s responses, and thereby inform how its offspring are generated. Secondly, when available, knowledge about the NN structure can also be leveraged to estimate the local effect of weight perturbations on an NN’s outputs[21].

3.4 Identification of somatic mutation without matching normal tissue

Predicts somatic mutations from tumor only samples. ISOWN is an algorithm that uses supervised machine learning to distinguish simple substitution somatic mutations in coding regions from germline variants in the absence of matching normal DNA. The software can assist researchers in accelerating sequencing process, reducing financial investment in sample sequencing and storing requirements, or increase the power of analysis by increasing the number of tumor samples sequenced with the same resources [4].

In ISOWN [4], at first they have downloaded, reformat and unzip external databases and annotated the validation cancer set using COSMIC v69, dbSNP v142, Mutation Assessor, ExAC r0.3, and PolyPhen-2. Then they gave facilities to the user to generate training data from user based on user data.

Some pre-processing such as filtering for passed filtered and some minimum read depth is done.

Furthermore, seven supervised machine learning algorithms have been applied such as JRip, J48,random forest, LADTree, naive Bayes classifier (NBC), Logistic regression, and Support vector machine (SVM) and uses external database for preprocessing the training data.

But Naive bayes algorithm has performed much better among the five cancer training datasets (COAD, UCEC, KIRC, BRCA, and PAAD). Here, LADTree and Random forest performed better in the remaining cancer training dataset (ESCA).

Their algorithm accurately classified between 95% to 98% of somatic mutations with F1-measure ranges from 75.9 to 98.6 depending on the type of cancer it is.

For PAAD dataset, it gives the worst F1-measure 75.9**.**

3.5 Summary

This chapter discusses about the previous works to identify and the ways to predict the mutation from DNA sequence along with some implementation of neural network in this.

Chapter 4

Proposed somatic mutation prediction approach

The proposed somatic mutation prediction approach is shown below:

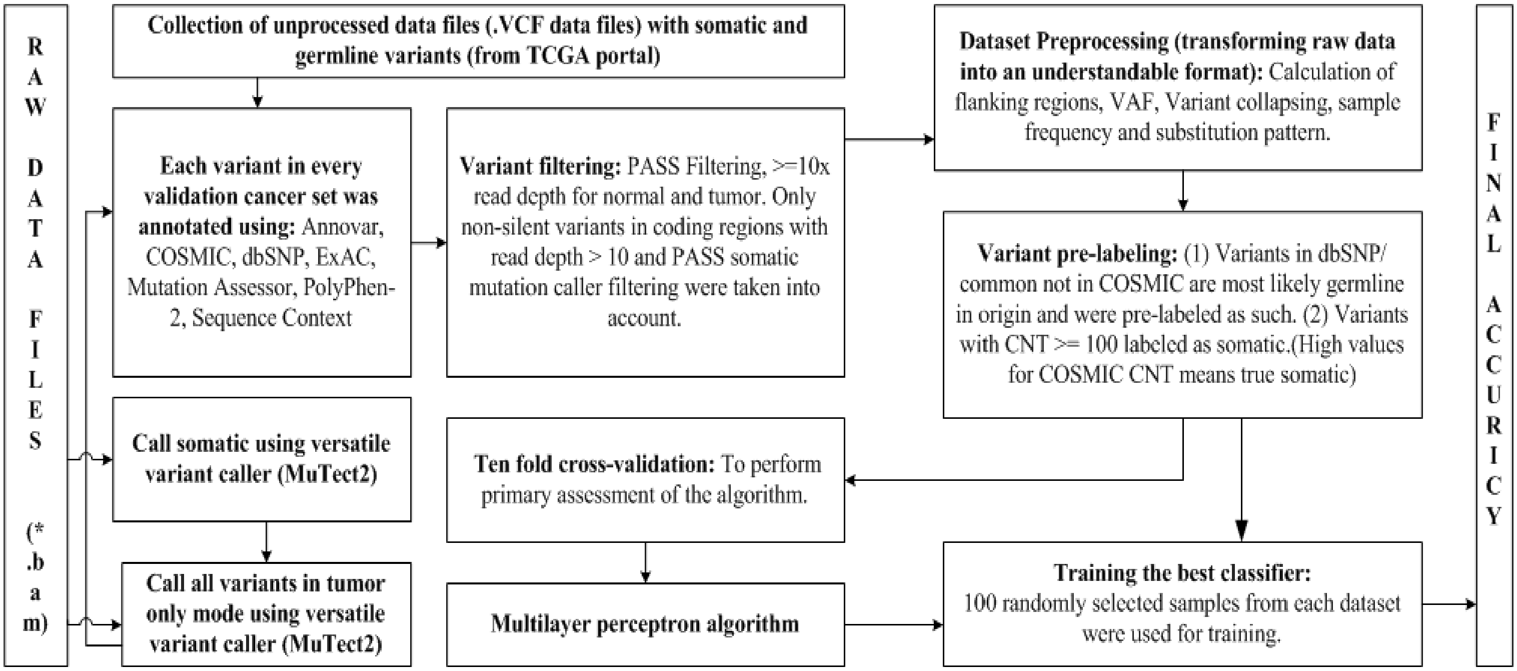

RAW DATA FILES (*.bam)
Collection of unprocessed data files (.VCF data files) with somatic and germline variants (from TCGA portal)
Each variant in every validation cancer set was annotated using: Annovar, COSMIC, dbSNP, ExAC, Mutation Assessor, PolyPhen-2, Sequence Context
Variant filtering: PASS Filtering, >=10x read depth for normal and tumor. Only non-silent variants in coding regions with read depth > 10 and PASS somatic mutation caller filtering were taken into account.
Call somatic using versatile variant caller (MuTect2)
Call all variants in tumor only mode using versatile variant caller (MuTect2)
Ten fold cross-validation: To perform primary assessment of the algorithm.
Multilayer perceptron algorithm
Dataset Preprocessing (transforming raw data into an understandable format): Calculation of flanking regions, VAF, Variant collapsing, sample frequency and substitution pattern.
Variant pre-labeling: (1) Variants in dbSNP/ common not in COSMIC are most likely germline in origin and were pre-labeled as such. (2) Variants with CNT >= 100 labeled as somatic.(High values for COSMIC CNT means true somatic)
Training the best classifier: 100 randomly selected samples from each dataset were used for training.
FINAL ACCURICY

Figure 1: Framework of Proposed System Architecture

4.1 Machine learning

The following measures were used to evaluate neural network performance:

a. ***Recall*** **(or sensitivity or true positive rate):**

It measures the proportion of the known somatic variants that are correctly predicted.

Mathematically, recall is defined as follows:

$TP/(TP + FN)$, where TP is true positive and FN is false negative.

b. ***Precision:*** Proportion of the true positive from cases that are predicted as positive.

Mathematically, Precision is defined as follows:

$TP/(TP+FP)$;where FP is false positive.

c. ***F1-measure:*** F1-measure is a measure of a test's accuracy.

It is the harmonic mean of precision and recall.

Mathematically, Precision is defined as follows:

$2 \times (\text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall})$.

d. ***False positive rate*** **(FPR):**

False positive rate is an outcome where the model *incorrectly* predicts the *positive* class.

In this case it measures the fraction of germline mutation which are incorrectly classified as somatic mutation.

Mathematically, false positive rate is defined as follows:

$FP/(FP+TN)$, where TN is true negative.

e. ***Accuracy*** **(ACC):**

Mathematically, Accuracy is defined as follows:

$(TP+TN)/(TP+FN+TN+FP)$.

f. ***False negative rate*** **(FN):** Suppose there is a cancer mutation that cannot be detected due to allele dropout during either single-cell genome amplification or sequencing, resulting in a false negative means wrongly classified as negative.

False negative rate is an outcome where the model *incorrectly* predicts the *negative* class.

In this case it measures the fraction of somatic mutation which are incorrectly classified as germline mutation.

Mathematically, false negative rate is defined as follows:

FNR = FN / (TP+FN)

4.2 Features Information of Training Data

There are 10 features in the training dataset. These features are described below -

4.2.1 The **Catalogue of Somatic Mutations In Cancer (COSMIC)**

COSMIC which means the Catalogue of Somatic Mutations in Cancer. It is the world's largest source of expert manually curated somatic mutation information relating to human cancers. Here, VCF files of both Coding and Non Coding mutations have been used. It is the richest database of the cancer related somatic mutation. As we have used both coding and non-coding mutations, one biggest drawback is that in coding somatic SNVs it is catalogued by COSMIC which were submitted from a single sample only. In which most of these are random mutations. So, therefore we have used COSMIC CNT attribute. CNT was used as a feature [10].

4.2.2 ExAc

The Exome Aggregation Consortium (ExAC) is a coalition of investigators seeking to aggregate and harmonize exome sequencing data from a wide variety of large-scale sequencing projects. ExAC has collection of 60,706 which has unrelated individuals sequenced as part of various disease-specific and population genetic studies and this is by far one of the richest databases of common germline polymorphisms [11].

4.2.3 dbSNP

dbSNP currently classifies nucleotide sequence variations with the following types and percentage composition of the database: (i) single nucleotide substitutions, 99.77%; (ii) small insertion/deletion polymorphisms, 0.21%; (iii) invariant regions of sequence, 0.02%; (iv) microsatellite repeats, 0.001%; (v) named variants, <0.001%; and (vi) uncharacterized heterozygous assays, <0.001% [7]. dbSNP contains both common germline variants and also rare polymorphisms. Here, validation sets were annotated against dbSNP/rare databases and dbSNP/common_all.

4.2.4 Sequence context

Sequence context is three-base sequence of human genome. It may vary from one cancer to another. But, there are certain pattern similarities between somatic mutation and germline mutation.

4.2.5 Mutation Assessor

Mutation assessor predicts the functional impact of amino-acid substitutions in proteins. Mutation assessor employs information based on the analysis of evolutionary conservation patterns in protein family multiple sequence alignments. It has been validated on a large set of disease associated and polymorphic variants. This tool enables the determination of mutations discovered in cancer or missense polymorphisms [12]. Its data type is categorical (high, medium, low, or neutral). Its value also can be Stop loss and stop gain which is annotated by annovar which has greater impact on protein function and can cause somatic alterations.

4.2.6 Polyphen (Polymorphism Phenotyping)

The Polymorphism Phenotyping (Polyphen) is a tool which predicts possible impact of an amino acid substitution on the structure and function of a human protein using straightforward physical and comparative considerations [16]. PolyPhen predicts the functional significance of an allele replacement from its individual features by a Naïve Bayes classifier. The web application allows users to (i) predict the effect of a single-residue substitution or reference single nucleotide polymorphism SNP, (ii) analyze SNPs in a batch mode, and (iii) search in a database of precomputed predictions for the whole human exome sequence space.

4.2.7 Sample frequency

Sample frequency indicates that sample in the particular dataset contains germline polymorphism or somatic polymorphism considering the total number of sample which is measured by fractional number. High value indicates germline polymorphism and low value indicates somatic polymorphism.

4.2.8 Variant allele frequency (VAF)

The allele frequency represents the incidence of a gene variant in a population. Alleles are variant forms of a gene that are located at the same position, or genetic locus, on a chromosome. An allele frequency is calculated by dividing the number of times the allele of interest is observed in a population by the total number of copies of all the alleles at that particular genetic locus in the population. Allele frequencies can be represented as a decimal, a percentage, or a fraction. In a population, allele frequencies are a reflection of genetic diversity. Changes in allele frequencies over time can indicate that genetic drift is occurring or that new mutations have been introduced into the population [15].

VAF for somatic mutation is mostly in between 22%-50% and germline mutation it is mostly centered at 50%.

But, due to copy number variation, admixture with normal tissues and/or tumor subclonality the value can differ.

4.2.9 Flanking region

Flanking region means that the DNA sequences is extended on either side of a gene. The work of flanking region is to measure whether a VAF of an unknown variant matches the VAF of a known germline polymorphism. Candidate variants were searched for flanking polymorphisms that is present within 2 Mbp of 5’ or 3’ flanking region and labeled as V1 and V2. Here, 5’ flanking region means it is not transcribed into RNA and also it has a complex set of regulatory elements for example enhancer, silencer etc. and here it is labeled as v1 [13]. On the other hand, 3’ flanking region is labeled as v2 and 3’ flanking region means it is discovered to be transcribed into RNA but in DNA.

4.2.10 Substitutional pattern

Substitution rates vary between species. There are six substitution subtypes. CA, CG, CT, TA, TC these are six categorical subtypes. Species with short generation time generally evolve faster, presumably because they experience more rounds of germ-cell divisions (and hence more DNA replication errors) during a given unit of time. If most mutations are due to DNA replication errors, then mutation rates are expected to be higher in males than in females. Also in autosomes, there are substantial variations in neutral substitution rates. Patterns of neutral substitution vary also at the gene scale [12].

Substitutional pattern is also the newly introduced variant base of mutation form. Such as if we consider chr3, 178936094C>G. Here, CG is the mutation. And as there are six subtypes of substitutional patterns are present. Therefore, this is one of them. In many datasets which were tested shows that in some substitution pattern, somatic mutation and germline are enriched [9].

Features	**Type Of value**	**Sample value**
COSMIC_CNT	Integer	0, 1
ExAC	Boolean	True, False
dbSNP	Boolean	True, False
Mutation assessor	Categorical	neutral,low,medium,high, stopgain,stop loss
PolyPhen-2	Categorical	benign, probably, possibly
Sequence context	Categorical	Three-base sequence like ATT,CTT,GTT etc.
Sample frequency (SF)	Double	In between 0 to 1
Variant allele frequency	Double	Depends
Flanking regions	Double	0 to 1, NA
Substitution Pattern	Categorical	CG, CA, CT, TA, TC, TG
isSomatic	Boolean	True, False

Table 1: Feature Information

4.3 Dataset generation

At first datasets were downloaded in VCF format from TCGA portal which contains both somatic variants and germline variants. Only from PAAD (TCGA-IB-7651-01A) was removed based on its highly mutational loads. In short, 300-fold in comparison to the median. Variant calling in KIRC (kidney renal clear cell carcinoma), PAAD (Pancreatic adenocarcinoma), and COAD (colon adenocarcinoma) sets was done using the Baylor College of Medicine (BCM) CARNAC (Consensus And Repeatable Novel Alterations in Cancer) pipeline (version 1.0); In BRCA (breast invasive carcinoma) and UCEC (uterine corpus endometrial carcinoma) sets with the bam bam pipeline (version 1.4) is from University of California at Santa Cruz [According to the headers of the retrieved VCF files. The KIRC, PAAD and COAD sets did not contain any homozygous variant. This can happen because of a consequence of CARNAC filtering. KIRC, PAAD and COAD sets did not contain any homozygous variants among the five TCGA datasets which were used for validation. As it is inconsistent so all homozygous variants from UCEC and BRCA have been removed to maintain the consistency across the five datasets. Other than that, 145 ESO (esophageal adenocarcinoma) BAM files from dbGAP portal have been downloaded and after that from the same files we extracted the raw reads. BWA (v 0.6.2) has been used to align them to human genome hg19. After that collapsed reads which aligned in a correct way were passed to the MuTect2 to call variants. Because, MuTect2 calls all

Variants in tumor only mode using versatile variant caller[9]. MuTect2 was run twice on two different modes. One is with pair matching normal and another one is using tumor_only_mode to call all somatic and germline variants. Pair matching normal which is basically a usual mode was used to retrieve gold-standard somatic mutation calls. And tumor_only_mode generally mimics the situation when these matching normal data are not available enough.For training set generation,variants from 100 ESO (esophageal adenocarcinoma) samples were selected randomly. And the other remaining samples were selected for the validation purposes. This training data generation steps were followed by previous work [4]. But, however our result do not depend upon the external databases for computation.

Finally, after the training dataset generation is completed, we applied neural network algorithm which is performed in 10 fold cross validation dataset.

4.4 Summary

This chapter shows the detail description of how the datasets were generated using the external genomic databases such as COSMIC, dbSNP, ExAC,PolyPhen, mutation assessor etc. Also, this chapter shows how we have calculated and measured the consequences using matching learning method.

Chapter 5

Role of artificial neural network in Human DNA sequence

Artificial neural network plays a vital role in human DNA exome sequence. Analysis of the DNA sequences of genes is important and necessary in the study of genetics. Many methods can be applied to their analysis, and the artificial neural network is one of the most frequent methods used to obtain the characteristic features of DNA sequences. Here we show how artificial neural networks acquire the features of DNA sequences.

5.1 A Brief discussion of Neural Network

Neural network in machine learning is a computing system which is inspired by the biological neural networks. It can be both the types supervised and unsupervised learning. This paper describes the implementation of neural network and in this experiment, we used supervised learning as outcome is known and given.

Neural network generally consist of three thing input layer, hidden layer and output layer. Input layer is consist of number of inputs and input layer is connected to another layer which is called hidden layer. As biological neural network concept, a neuron fires in certain threshold, it is also applied in here also. There are some weights normally initialized randomly in each layer neuron to other layer neuron and for certain threshold which is calculated by activation function like sigmoid, relu, softmax etc. a neuron becomes active.

Hidden layer can be any number but not much that may cause slowdown of computation. Neural network consist of large hidden layer called deep neural network. In neural network, there is a term called backpropagation after feedforwording to the one layer to another till the output layer. It backpropagate the network and calculate error and updates weight so that it can give better result. There is also a term in neural network called learning rate indicates how much the model learns from the network. Value should be picked precisely. To avoid overfitting, dropout is used that randomly drops some connection between neurons. Neural network give good outcome after time to time training much data.

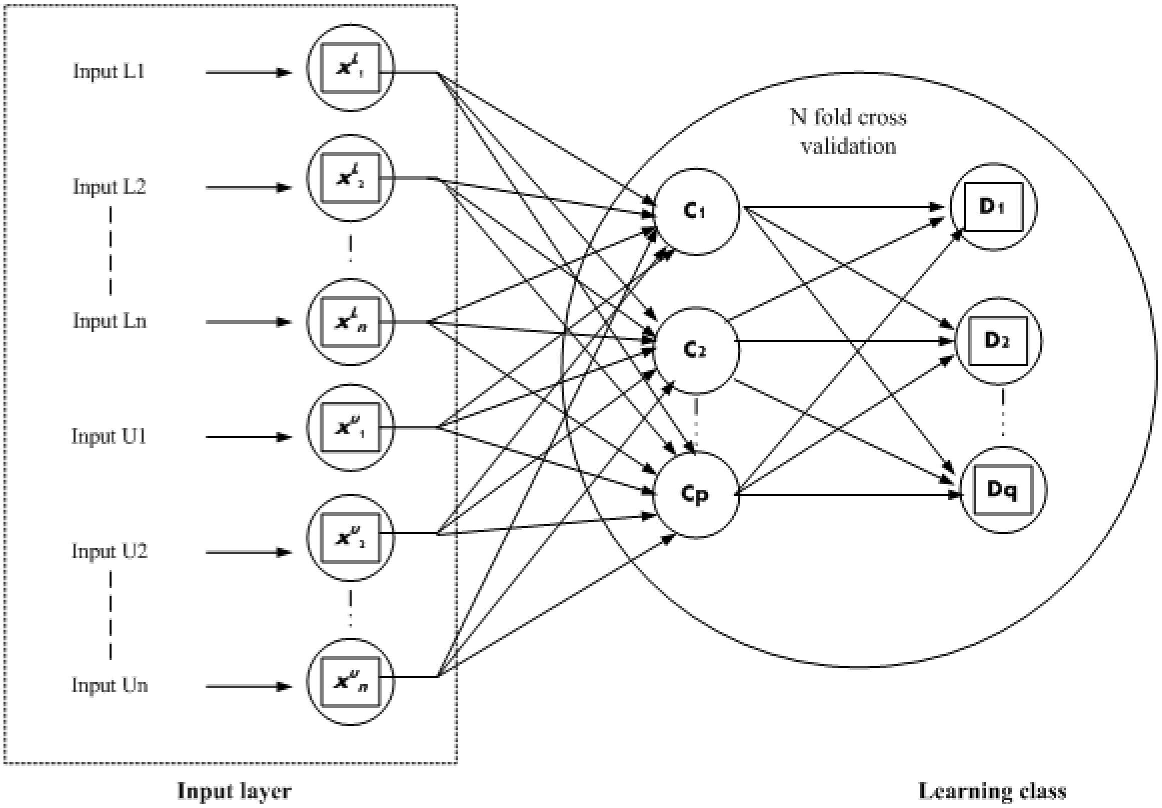


Figure 2: A simple neural network

5.2 Neural network for genomics

Some of the first applications of neural networks in genomics involved training single-layer fully connected neural networks on gene expression data, typically after using principal component analysis to reduce the dimensions of the input. These networks were used to distinguish between tumor types, predict tumor grade, and predict patient survival from gene expression patterns. Improvements to this included developing feature selection techniques for neural networks that could identify subsets of genes or (signatures) that were most predictive[25].

5.3 Cross Validation

Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample.

The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation. When a specific value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation.

Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model.

It is a popular method because it is simple to understand and because it generally results in a less biased or less optimistic estimate of the model skill than other methods, such as a simple train/test split.

The general procedure is as follows:

1. Shuffle the dataset randomly.
2. Split the dataset into k groups
3. For each unique group:
    1. Take the group as a hold out or test data set
    2. Take the remaining groups as a training data set
    3. Fit a model on the training set and evaluate it on the test set
    4. Retain the evaluation score and discard the model
4. Summarize the skill of the model using the sample of model evaluation scores

Importantly, each observation in the data sample is assigned to an individual group and stays in that group for the duration of the procedure. This means that each sample is given the opportunity to be used in the hold out set 1 time and used to train the model k-1 times.

It is also important that any preparation of the data prior to fitting the model occur on the CV-assigned training dataset within the loop rather than on the broader data set. This also applies to any tuning of hyperparameters. A failure to perform these operations within the loop may result in data leakage and an optimistic estimate of the model skill.

The results of a k-fold cross-validation run are often summarized with the mean of the model skill scores. It is also good practice to include a measure of the variance of the skill scores, such as the standard deviation or standard error [26].

5.4 Summary

This chapter shows the description of how neural network works in general and how neural network works for genomic data. Also this chapter describes the cross-validation and its aspects.

Chapter 6

Performance of Proposed Method

Finally, after successful completion of data generation its time to go to the next step. In the next step, we have applied a machine learning algorithm. In this work, we used artificial neural network. The algorithm is briefly discussed in 5.1.

After applying neural network algorithm in the dataset it's time to check out the performance.

6.1 Performance on training dataset

There are six training dataset. To measure how neural network performs we first applied it in the training dataset. It gives us an idea how much this algorithm learns in the training dataset.

Training data name	No of Hidden layers	Learning Rate	Accuracy	F1-measure	Recall	Precision	False positive rate
BRCA	5	0.01	96.5295%	0.965	0.965	0.965	0.035
COAD	2	0.05	95.9032%	0.959	0.959	0.959	0.041
ESCA	3	0.02	88.9342%	0.889	0.889	0.889	0.111
KIRC	5	0.01	97.0498%	0.970	0.970	0.971	0.030

Training data name	No of Hidden layers	Learning Rate	Accuracy	F1-measure	Recall	Precision	False positive rate
PAAD	5	0.01	96.1651%	0.962	0.962	0.962	0.038
UCEC	5	0.01	98.0983%	0.981	0.981	0.981	0.019

Table 2: Performance on training data

6.2 Performance on 10 fold cross validation

Applying neural network in only training dataset is not enough to measure how good this algorithm is. We could have used test data from training data by splitting it. Let’s say, we could have split the training data to 10 or 20% for testing. But, however K-fold cross validation in our case 10 fold cross validation gives better idea about an algorithm performance. Justification of choosing 10 fold cross validation is briefly discussed in 5.3

Training data name	No of Hidden layers	Learning Rate	Accuracy	F1-measure	Recall	Precision	False positive rate
BRCA	5	0.01	94.1889%	0.942	0.942	0.942	0.058
COAD	2	0.05	94.9463%	0.949	0.949	0.949	0.051
ESCA	3	0.02	88.0475%	0.880	0.880	0.881	0.120
KIRC	5	0.01	94.7264%	0.947	0.947	0.947	0.053

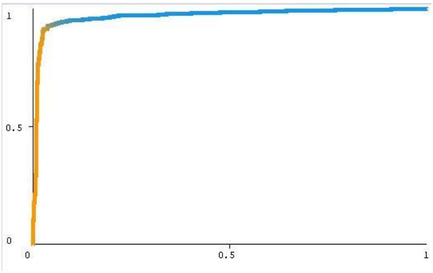
Training data name	No of Hidden layers	Learning Rate	Accuracy	F1-measure	Recall	Precision	False positive rate
PAAD	5	0.01	92.8425%	0.928	0.928	0.928	0.072
UCEC	5	0.01	97.277%	0.973	0.973	0.973	0.027

Table 3: Performance on 10 fold cross validation

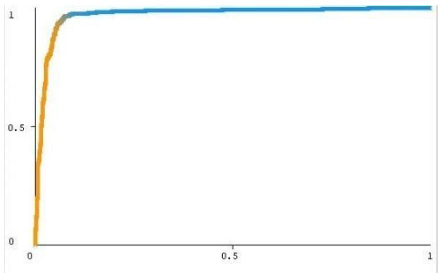
We used learning rate of 0.01 and hidden layer of 5 except for COAD and ESCA as the data is big and to minimize time. Neural network's computation time depends on the number of hidden layer and learning rate.

Area Under the ROC Curve:

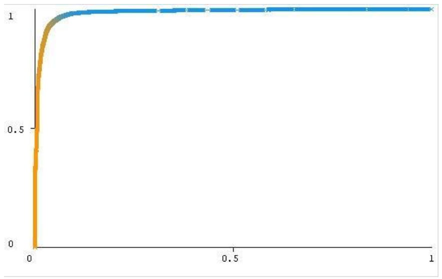
AUC stands for "Area under the ROC Curve." That is, AUC measures the entire two-dimensional area underneath the entire ROC curve (think integral calculus) from (0,0) to (1,1).
AUC provides an aggregate measure of performance across all possible classification thresholds. One way of interpreting AUC is as the probability that the model ranks a random positive example more highly than a random negative example.
The Area Under the ROC Curve for tested true and false case for both training data and 10 fold cross validation on all six cancer dataset is shown below:



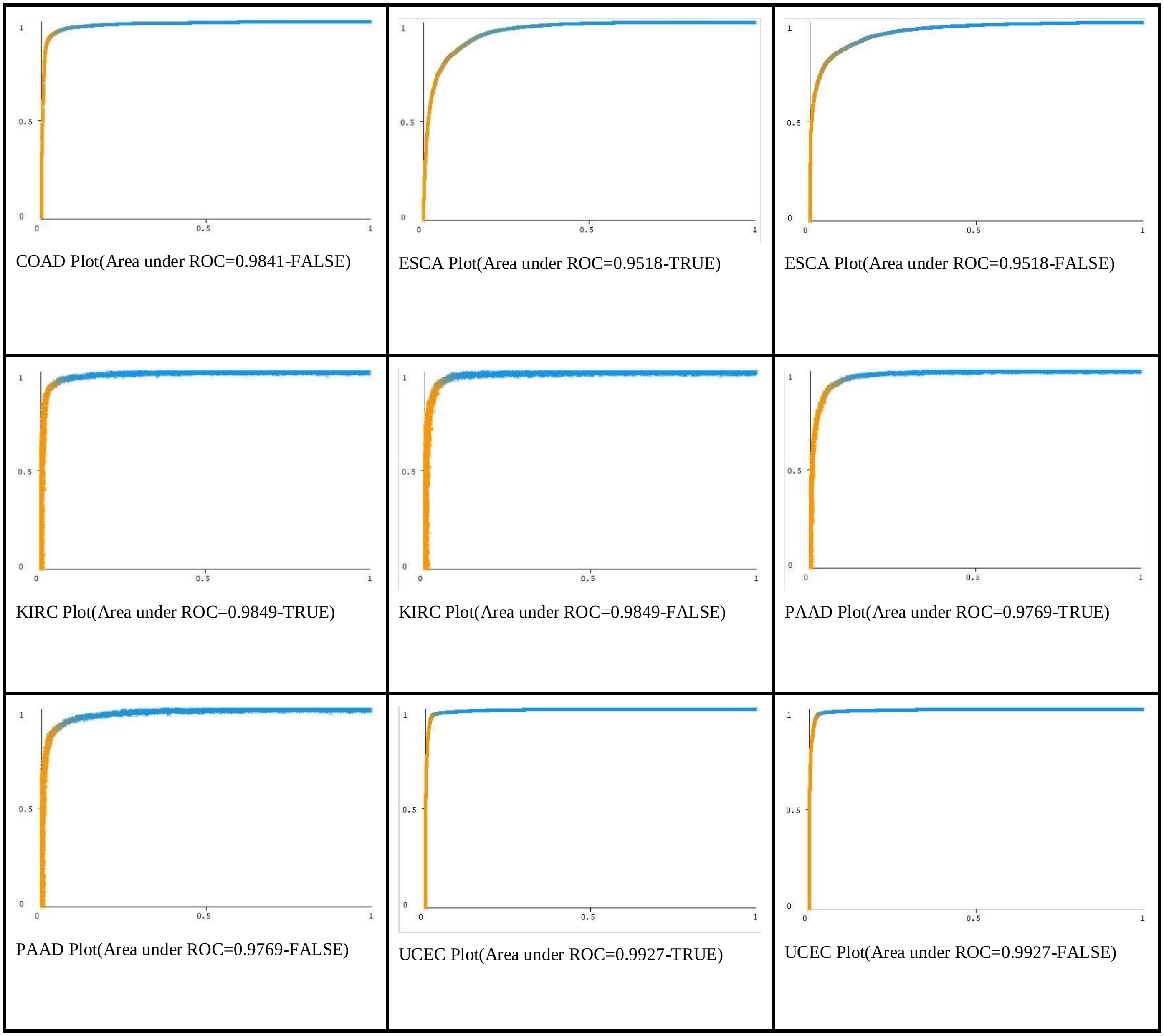
BRCA Plot(Area under ROC=0.9683-TRUE)



BRCA Plot(Area under ROC=0.9683-FALSE)



COAD Plot(Area under ROC=0.9841-TRUE)



COAD Plot(Area under ROC=0.9841-FALSE)

ESCA Plot(Area under ROC=0.9518-TRUE)

ESCA Plot(Area under ROC=0.9518-FALSE)

KIRC Plot(Area under ROC=0.9849-TRUE)

KIRC Plot(Area under ROC=0.9849-FALSE)

PAAD Plot(Area under ROC=0.9769-TRUE)

PAAD Plot(Area under ROC=0.9769-FALSE)

UCEC Plot(Area under ROC=0.9927-TRUE)

UCEC Plot(Area under ROC=0.9927-FALSE)

Figure 3: Change of Plot of AUC for different type of cancer

In Fig. 3 AUC curve is displayed for six different exome sequencing cancer sets(data) where some of them are for tested true and some others are for tested false.

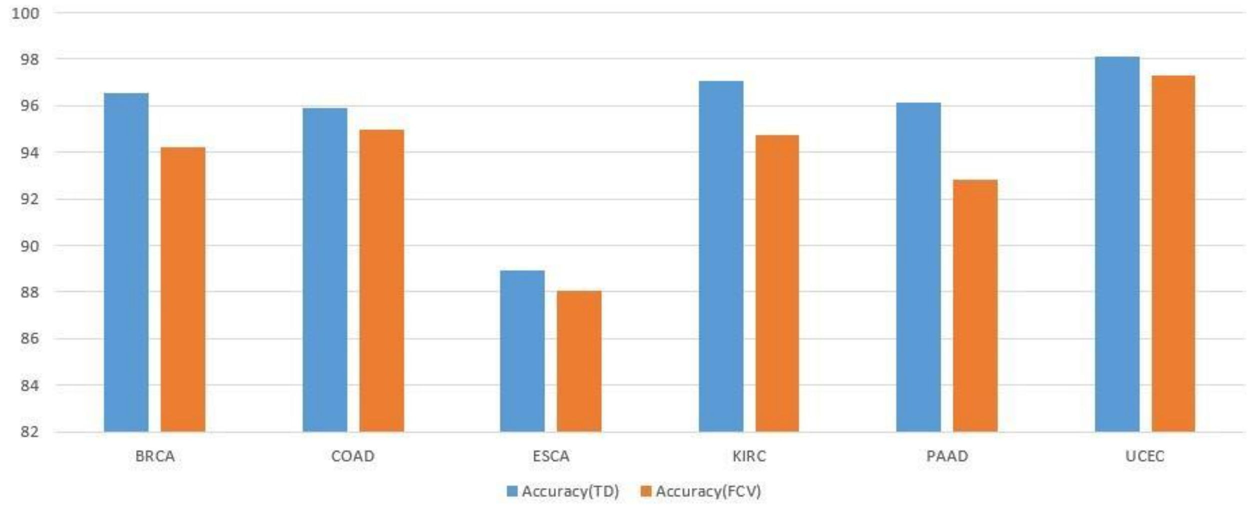


Figure 4: Comparison of Accuracy Between Training data and 10 Fold Cross Validation

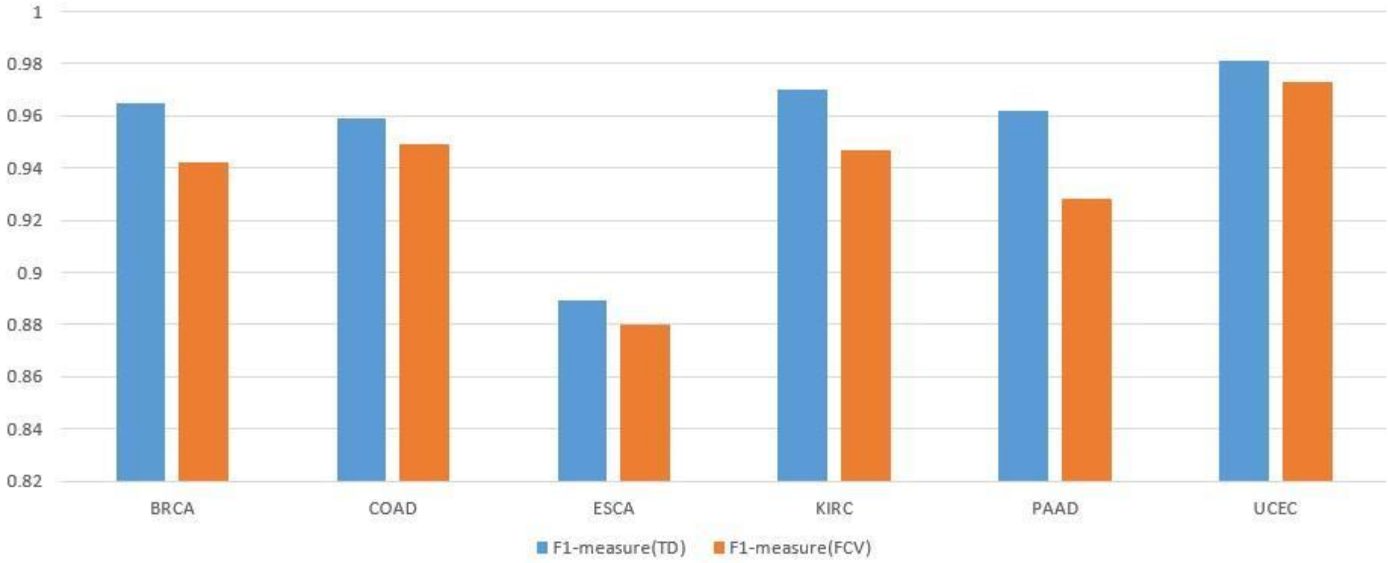


Figure 5: Comparison of F1-Measure Between training data and 10 Fold Cross Validation

In figure 4 and figure 5 demonstrates the comparison of accuracy and f1-measure between training data and 10 fold cross validation.

It is quite understandable that the performance on 10 fold cross validation is lower than the performance on training data. Though the deviation of the performance is not so high that indicates that neural network doing quite well in terms of learning and predicting.

6.3 Implementation of Python:

Now, it’s time for some coding. The coding is almost same except the loading the different dataset. So, here we are giving the coding for BRCA dataset.

Loading data:

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
df = pd.read_csv('BRCA_100_TrainSet.csv', names=["inExAct", "dbSNP", "CNT", "fre", "VAF","mutAss", "pattern","SeqContent","isFlanking","polyphen","isSomatic","tobedeleted"])
```

Deleting Unnecessary column:

```
df.drop('tobedeleted', axis=1, inplace=True)
```

This column is unnecessary. It arises due to file format.

Checking the data types:

For any other preprocessing purposes, let’s check the data type first.

```
df.dtypes
```

Converting data from Categorical to Nominal:

There are some categorical data in the dataset. As we know neural network don’t support categorical data so we have convert it.

There are categorical data in mutAss, pattern, Seq Content and polyphen. We convert it nominal data which stats from 1.

Example:

```
df.replace('benign',1,inplace=True)
df.replace('probably',2,inplace=True)

df.replace('neutral',1,inplace=True)
df.replace('low',2,inplace=True)
df.replace('CG',1,inplace=True)
df.replace('CA',2,inplace=True)

df.replace('ATT',1,inplace=True)
df.replace('CTT',2,inplace=True)
```

Missing Data:

There are some missing data in the dataset. So, we first replace this data from “?” to np.nan. Then, we take the mean of the column to fill the missing data.

```
df.replace('?',np.nan,inplace=True)
df.fillna(df.mean(), inplace=True)
```

Converting Data-types:

There are boolean values in inExAct, dbSNP and isSomatic. So, we convert it to integer. And, In isFlanking the given data is float but it shows string so we convert it to float.

```
df['inExAct'] = df.inExAct.astype(int)
df['isFlanking'] = df.isFlanking.astype(float)
df['dbSNP'] = df.dbSNP.astype(int)
df['isSomatic'] = df.isSomatic.astype(int)
```

Defining target variable:

At this point, we define the target variable (output). In this case, it is isSomatic.

```
X = df.iloc[:, 0:10].values
y = df.iloc[:, 10].values
```

Building a Neural network model and evaluating training dataset:

```
model = Sequential()
model.add(Dense(15, input_dim=10, activation='relu'))
model.add(Dense(10, activation='relu'))
model.add(Dense(10, activation='relu'))
model.add(Dense(10, activation='relu'))
model.add(Dense(10, activation='relu'))
model.add(Dense(10, activation='relu'))
model.add(Dense(1, activation='sigmoid'))
# Compile model
model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])
# Fit the model
model.fit(X, y, epochs=150, batch_size=10,verbose=0)
# evaluate the model
scores = model.evaluate(X, y)
print("\n%s: %.2f%%" % (model.metrics_names[1], scores[1]*100))
```

Building a Neural network model and evaluating 10 Fold Cross Validation:

```
from sklearn.model_selection import StratifiedKFold
seed = 7
numpy.random.seed(seed)
kfold = StratifiedKFold(n_splits=10, shuffle=True, random_state=seed)
cvscores = []
for train, test in kfold.split(X, y):
    model = Sequential()
    model.add(Dense(15, input_dim=10, activation='relu'))
    model.add(Dense(10, activation='relu'))
    model.add(Dense(10, activation='relu'))
    model.add(Dense(10, activation='relu'))
    model.add(Dense(10, activation='relu'))
    model.add(Dense(10, activation='relu'))
    model.add(Dense(1, activation='sigmoid'))
    # Compile model
    model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])
    # Fit the model
    model.fit(X[train], y[train], epochs=150, batch_size=10, verbose=0)
    # evaluate the model
    scores = model.evaluate(X[test], y[test], verbose=0)
    print("%s: %.2f%%" % (model.metrics_names[1], scores[1]*100))
    cvscores.append(scores[1] * 100)
print("%.2f%% (+/- %.2f%%)" % (numpy.mean(cvscores), numpy.std(cvscores)))
```

6.4 Decision tree

We have generated decision tree for KIRC,PAAD,UCEC,ESCA and COAD. But as they are so long (tree) and these are so difficult to see here. For better understanding BRCA is shown here. By applying j48 we get the decision tree of BRCA which is shown below:

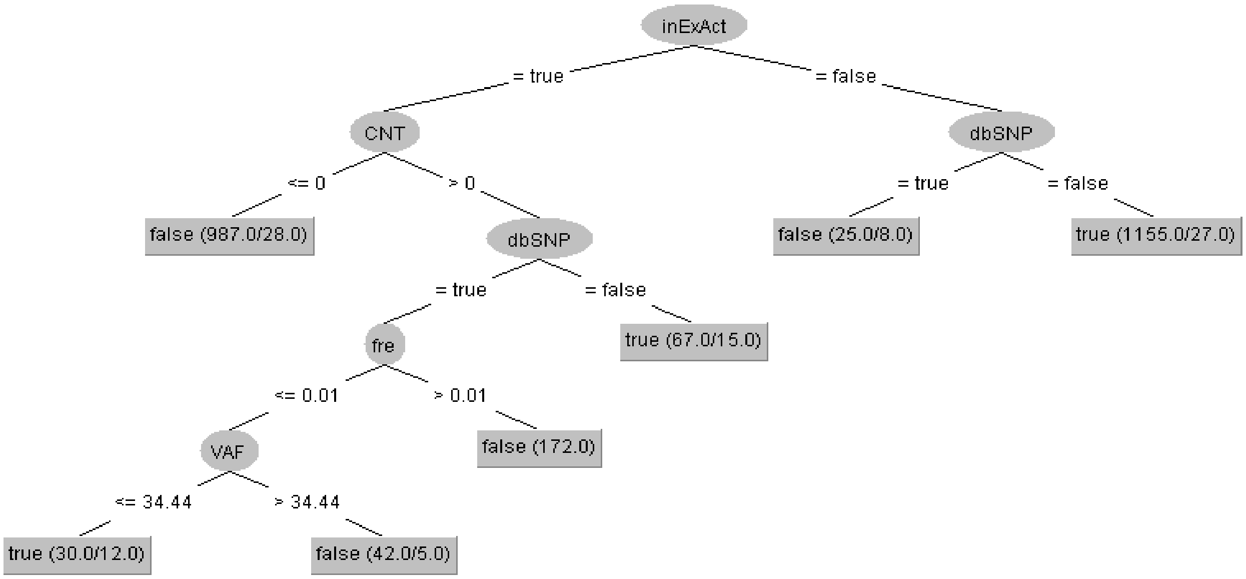


Figure 6: Decision tree of BRCA

6.5 Evaluation of Classifier Performance

In this work, an efficient approach is proposed which can perform on the training datasets which was generated after annotating, filtering, preprocessing and pre-labeling the external databases and making a testing sample set (See Fig.1). The training sets were given as well as the training set have generated also. In both cases we have found the same result for the datasets. In this proposed method, the algorithm can work without using the testing sample set and the overall architecture of this pipeline can be summarized by Fig. 2.

This efficient method performed better without using the testing predicted file came from the external databases like COSMIC, ExAC, dbSNP, Mutation Assessor and PolyPhen. The Performance can be increased to a high level by removing the missing values from the external databases or by filling up the fields of the missing value to some extent. The performance of the machine learning classifiers was evaluated using tenfold cross-validation based on the training datasets. The performance was checked for each of the six cancer type data sets where each balanced with 700 somatic and 700 germline variants which was preprocessed and collapsed independently and the accuracy was calculated using neural network algorithm. BRCA without polyphen correctly classified 94.0678% and without flanking and with polyphen it correctly classified 94.1082% also f-measure of BRCA is 0.942 using multi layer perceptron algorithm and it classified correctly 89.7498%. In COAD dataset, using multilayer perceptron algorithm in 10 fold cross validation and the accuracy was 94.9463% with all the flanking and polyphen, having 0.051 false positive rate. ESC using multilayer perceptron, correctly classified 88.0475% with the highest number of false positive rate while without flanking it gives 88.0194% but without polyphen it was 88.1179%. With multilayer perceptron the accuracy of KIRC was 94.7264% having the same f-measure, recall and precision of 0.947. But without isflanking it gives 94.6194%. The accuracy of PAAD using multilayer perceptron was 92.8425%.

But if we remove flanking and polyphen using the same algorithm (multilayer perceptron) still it performs better (93.2377%). Now, UCEC has the less number of false positive rate 0.027 also has the same 0.973 rate in f-measure, recall and precision having the highest accuracy of 97.277%. UCEC performed much better with neural network than any other.Also, if we remove flanking and polyphen, still it performs the highest accuracy which is 97.2054%. If we remove polyphen using the multilayer perceptron algorithm, still this is not higher than the actual accuracy. So, overall the accuracy of UCEC was best and ESCA performed worst accuracy.

6.6 Analyze Feature Factors

Furthermore, to estimate whether a certain feature can cause somatic mutation or not we have applied Decision tree (J48) to visualize and analyze the feature factors that are responsible for somatic mutation or not. And some meaningful information has been found through this.

In BRCA dataset, almost in about 40% cases, if inExAct is true and CNT <= 0 then it's most likely to be not somatic. On the other hand, in 46% cases, if inExAct is false and dbSNP is false then it's most likely to be somatic.

In COAD dataset, we found out in about 38% cases, if CNT <= 0 and inExAct is true then it's most likely to be not somatic. In about 41% cases, if CNT > 0, fre <= 0.01 and dbSNP is false then it's most likely to be somatic. 6%cases shows, if CNT > 0, fre <= 0.01, dbSNP is true, VAF <= 41.565 and pattern is CT 2995 then it's most likely to be somatic.

In ESCA dataset, we found out in about 17% cases, if dbSNP is true, VAF > 31.82, fre <= 0.01 and inExAct is true then it's most likely to be not somatic. 11% cases shows that if dbSNP is true, VAF > 31.82 and fre > 0.01 then it's most likely to be not somatic. In about 36% cases, if dbSNP is false, VAF > 3.976, fre <= 0.02 and inExAct is false then it's most likely to be somatic.

In KIRC dataset, we found out in about 33% cases, if inExAct is true, VAF > 36.13, CNT <= 0 then it's most likely to be not somatic. In about 40% cases, if inExAct is false, fre <= 0.01, CNT > 0 then it's most likely to be somatic.

In PAAD dataset, we found out in about 42% cases, if VAF > 34.92, inExAct is true then it's most likely to be not somatic. In about 31% cases, if VAF <= 34.92, fre <= 0.01, inExAct is false then it's most likely to be somatic.

In UCEC dataset, if inExAct is true, CNT <= 0 and VAF > 22.9 then it's most likely to be not somatic and we have found out this in about 40% cases. Moreover, 43% cases shows, if inExAct is false, dbSNP is false, CNT > 0 then it's most likely to be somatic.

6.7 Discussion

In previous work [4], predicted file from the external databases and the performance of somatic mutation and germline mutation were checked using tenfold cross validation but in this method, we have used multilayer perceptron without using the predicted test file and this process shows how algorithm worked well in many cases without the predicted files. We have received F1-measure ranging from 88 to 97% across multiple tumor types. If the missing values of polyphen were removed or if the missing values of polyphen can be replaced by something which can be used to identify the values,than it would definitely performed better. One of the main challenges was to replace those missing values as it is a large database which contains more than 50% of missing value in polyphen. Assessment of neural network in somatic mutation shows that you don’t need to rely on or don’t have to be dependent upon the test predicted files which was generated by the input directory which contains the input files in VCF format.

6.8 Summary

This Chapter shows how the performance on training datasets along with how the performance on 10 fold cross validation was. The Area under ROC curve and also the other comparisons have shown in graph to visualize the differences that we get after the comparison. How the multilayer perceptron algorithm was used and worked well. This chapter also describes what will happen if some of the feature were introduced and some of the feature were removed.

Chapter 7

Conclusions

7.1 Conclusions

This paper describes an efficient approach to distinguish the somatic single nucleotide variants in absence of matching normal tissue samples.

The proposed model uses machine learning along with external database and implemented across six different cancer types.

In this experiment, we achieved up to 97% accuracy with f1-measure range from 88 to 97% which is higher than other state-of-the-art methods .

Our approach is useful when normal tissue sample is not available and our result is not depend on external database so if training data is already generated then we don’t need no other things for computation thus it can reduce computation cost and financial (storage) cost.

7.2 Future Work

An interesting future work can be the improvement of its classifier performance as well as introducing more ways to enhance the capability to identify the somatic mutations accurately and the plan is to make more additional training sets through this approach and also finding more rules as well as additional features or characteristics through preprocessing the data.

Another future work can be done which is make a system that helps to generate dataset if there is slightest change in the database that we depend upon to generate training data.

References

[1] U.S. National Library of Medicine, National Institutes of Health, Department of Health & Human Services, 2018, July 17, Help Me Understand Genetics Mutations and Health, Retrieved from https://ghr.nlm.nih.gov/

[2] Amar, D, Izraeli, S, Shamir, R,"Utilizing somatic mutation data from numerous studies for cancer research: proof of concept and applications"Oncogene Volume 36, pages 3375–3383 (15 June 2017)

[3] Campbell, Catarina D,Eichler, Evan E,"Properties and rates of germline mutations in humans",Published on 2013 May 16 ,Genet. 2013 Oct; 29(10): 575–584

[4] Kalatskaya, Irina, Trinh, Quang M., Spears,Melanie, McPherson, John D., Bartlett, John M. S.,and, Stein, Lincoln, "accurate somatic mutation identification in the absence of normal tissue controls", *Genome Medicine 20179*:59, 29 June 2017

[5] H, Shen, J, Li, J, Zhang, C, Xu, Y, Jiang, Z, Wu, F, Zhao, L, Liao, J, Chen, Y, Lin, Q, Tian, CJ, Papasian, HW, Deng, "Comprehensive characterization of human genome variation by high coverage whole-genome sequencing of forty four Caucasians", Center for Bioinformatics and Genomics, Department of Biostatistics and Bioinformatics, School of Public Health and Tropical Medicine, Tulane University, New Orleans, Louisiana, USA., 2013 Apr 5

[6] Jauregui, Gonzaga, JR, Lupski, RA, Gibbs,"Human genome sequencing in health and disease", Department of Molecular and Human Genetics, Baylor College of Medicine.

[7] S., T. Sherry, M.-H., Ward, M., Kholodov, J., Baker, L., Phan, E., M. Śmigielski, K., Sirotkin, "dbSNP: the NCBI database of genetic variation",

Nucleic Acids Research, Volume 29, Issue 1, 1 January 2001, Pages 308–311, 01 January 2001

[8] Gonzaga-Jauregui, Claudia, Lupski, James R., Gibbs, Richard A, "Human Genome Sequencing in Health and Disease". Vol. 63:35-61, February 2012

[9] Cai, Lei, Yuan, Wei, Zhang, Zhou, He, Lin, Chou, Kuo-Chen, ”In-depth comparison of somatic point mutation callers based on different tumor next-generation sequencing depth data”, 22 November 2016

[10] Forbes, SA, Bhamra, G, Bamford, S, Dawson, E, Kok, C, Clements, J, Menzies, A, Teague, JW, Futreal, PA, Stratton, MR.,“The Catalogue of Somatic Mutations in Cancer (COSMIC)”, Wellcome Trust Genome Campus, Hinxton, United Kingdom, Chapter 10:Unit 10.11, 2008 Apr

[11] McKee H., Zheng K., Chan G., Ho J, “ExAC: Revolutionizing Rare Disease Diagnosis.”, ExAC: Revolutionizing Rare Disease Diagnosis. Illustrated by M. Yi. Rare Disease Review, January 2017.

[12] Duret L (2009), “Mutation Patterns in the Human Genome: More Variable Than Expected”, PLoS Biol 7(2): e1000028, February 3, 2009.

[13] Sabui, Subrata, Ghosal, Abhisek, Said, Hamid M., “Identification and characterization of 5′-flanking region of the human riboflavin transporter 1 gene (*SLC52A1*)”, Department of Medicine and Physiology/Biophysics, University of California-Irvine, Irvine, CA 92697, USA; Department of Medical Research, Veterans Affairs Medical Center, Long Beach, CA 90822, USA, 2014 Oct 5

[14] Dertat, A., “Applied Deep Learning - Part 1: Artificial Neural Networks”,20 July 2018, Retrieved from https://towardsdatascience.com/

[15] Sallman, David A., Padron, Eric,“Integrating mutation variant allele frequency into clinical practice in myeloid malignancies.”, Department of Malignant Hematology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL, USA. Hematology/Oncology and Stem Cell Therapy.Volume 9, Issue 3, September 2016, Pages 89-95

[16] Adzhubei, Ivan, Jordan, Daniel M.,Sunyaev, Shamil R., “Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2”, 2015 Jun 25

[17] Fan Y, Xi L, Hughes DS, Zhang J, Zhang J, Futreal PA, Wheeler DA, Wang W. MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. Genome Biol. 2016;17:178.

[18] Liu Y, Loewer M, Aluru S, Schmidt B. SNV Sniffer: an integrated caller for germline and somatic single-nucleotide and indel mutations. BMC Syst Biol. 2016;10 Suppl 2:47.

[19] Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics. 2012;28:1811–7.

[20] Hofree, Matan, Carter, Hannah, Kreisberg, Jason F., Bandyopadhyay, Sourav, Mischel, Paul S., Friend, Stephen, Ideker, Trey, Challenges in identifying cancer genes by analysis of exome sequencing data, Nature Communications volume7, Article Number: 12096(2016), 15 July 2016

[21] Lehman, Joel, Chen, Jay, Clune, Jeff, Stanley, Kenneth O., Safe Mutations for Deep and Recurrent Neural Networks Through Output Gradients, GECCO '18, July 15–19, 2018, Kyoto, Japan

[22] Loewe, L. (2008), Genetic mutation, *Nature Education* 1(1):113 Retrieved from https://www.nature.com/scitable/topicpage/genetic-mutation-1127

[23] Biology Wise Staff, July 14, 2017, A Helpful Guide to Understanding Somatic Mutation With Examples, Retrieved from https://biologywise.com/

[24] Single Nucleotide Polymorphism(SNP) | Learn Science at Scitable, 10 November,2015, Retrieved from *www.nature.com*

[25] Rao, Pooja, Neural Networks for Genomics, September 2, 2016, Retrieved from http://blog.qure.ai/

[26] Brownlee, Jason, Statistics Archive, A Gentle Introduction to k-fold Cross-Validation May 23, 2018, Retrieved from https://machinelearningmastery.com