Ain Shams University
Faculty of Computer and Information Sciences
BIOINFORMATICS PROGRAM

Detection of pre-cancer cells

By:

Mennatullah Fathy Sayed Hussien. [BioInformatics]
Rehab Essam Abdelghany. [BioInformatics]
Mohaned Kamal Mohamed. [BioInformatics]
Haidy Said Tawfik. [BioInformatics]
Nadeen Mostafa Abd-elsayed. [BioInformatics]
Manar El Sayed Abd El Meged. [BioInformatics]

Under Supervision of:

Mahmoud Mounir
Assistant Professor,
Information system department,
Faculty of Computer and Information Sciences,
Ain Shams University

Walaa Samir
Teaching Assistant,
Bioinformatics department,
Faculty of Computer and Information Sciences,
Ain Shams University

Acknowledgement

We would like to thank Dr. Mahmoud Mounir for pushing us to be our best selves, helping us in completing this project and supporting us throughout the whole year. We would like to express our gratitude for everyone who helped us during the whole years of study starting with endless thanks for our families in encouraging us to do a great job, our friends who are always support us, our DR's and TA's for teaching us across 4 years of education in our college (Faculty of Computer and Information Science).

Abstract

Mutations in the TP53 gene are the most commonly acquired mutations in cancer and pre-cancer. The p53 protein, made by the TP53 gene, normally acts as the supervisor in the cell as the body tries to repair damaged DNA. Different mutations can determine how well or how poorly that supervisor is able to direct the response .

Detecting pre-cancer stage is based on searching for mutations in gene sequences that cause cancer, somatic mutations in P53 protein that causes Pre-Cancer is still a challenge in cancer analysis, Due to mutations that are occurring in the TP53 gene to the cells inside the updated UMD TP53 Mutation Database as of October 2017, the tumor suppressor P53 is approximately approaching 50% of all human malignancies, The defect in function of the protein P53 is one of the common genetic alterations in human cancer.

Using of preprocessing steps and to construct the prediction model by selecting (35) out of (132) new TP53 gene database fields in order to classify the cases to the target Class pathology (Cancer, Pre-cancer) using these fields.

In this work, an artificial neural network is presented to predict cancer and pre-cancer caused by specific 2 codons mutations of the tumor protein P53 (each codon has hundreds of mutations that cause tumors). This method was applied on mutability of Codons 248 & 249, the used algorithm is very accurate in term of accuracy (Train data= 95.7%, Test data=94%)

Table of Contents

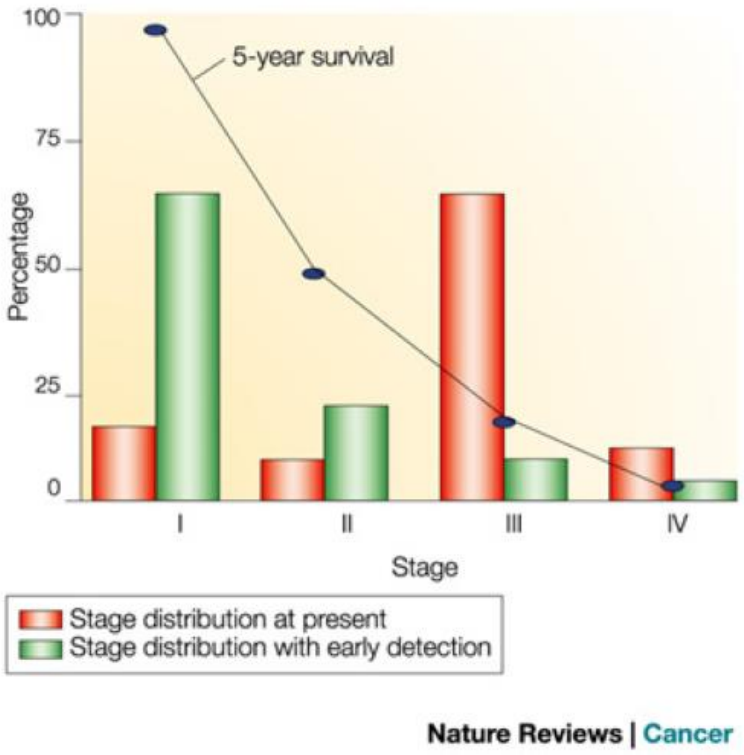
List Of Figures

Chapter 1

Introduction

1.1 Motivation

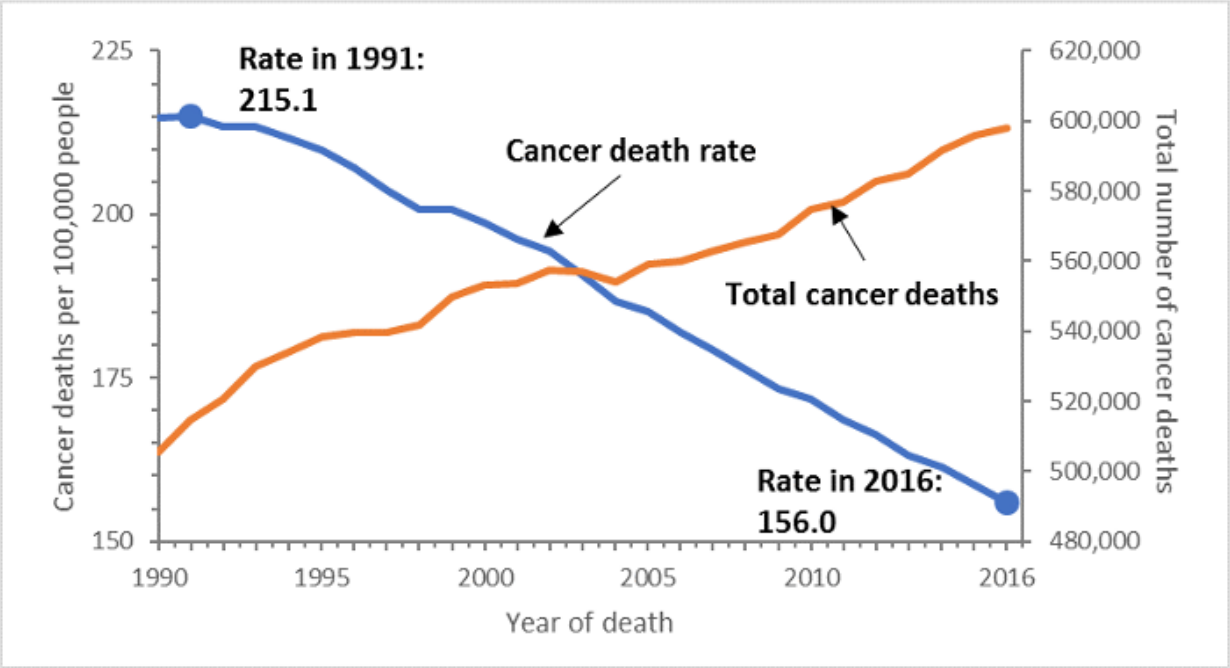
Cancer refers to one of a large number of diseases characterized by the development of abnormal cells that divide uncontrollably and have the ability to infiltrate and destroy normal body tissue. It often has the ability to spread throughout the body. It is the world's second largest cause of death. It screening aids in the early detection of cancer, when it is easier to treat. Early detection may result in shorter therapy and recovery time. The earlier you are diagnosed with cancer, the greater your chances of survival are, so detecting pre-cancer cell is very important factor to prevent cancer from forming and spreading too much as shown in Fig. [1.1].



(Fig1.1) number of people diagnosed with cancer decreases as the time for which the disease is detected increases.

1.2 Problem Definition

- Cancer deaths and the causes of discover it in final stage.
- Statistically, the death toll due to this disease has increased drastically as shown in Fig[1.2].



(Fig 1.2) Cancer mortality rate from 1991 to 2016

1.3 Objective

- Create a desktop-Application to allow the patient to enter his data and give him the result easily.
- Detecting mutations that cause (pre-cancer, cancer) cell.
- Reducing the number of cancer cases in the early stages of the disease.
- To identify and catalogue the underlying patterns of mutation that give rise to many different cancer types like (Wilms tumor, Breast tumor, Skin tumor, prostate cancer, etc.).
- Early diagnosis may aid in the creation of a patient-specific treatment plan.

1.4 Document Organization

❖ Chapter 2: Background

This Chapter discusses the field of the project and the biological and scientific basis on which the project is built, and mentions similar systems to our project and it finally discusses the technologies and algorithms used in the project.

❖ Chapter 3: Analysis and Design

This Chapter show the Architecture of the Project and some other diagrams and how the patient will interact with the system to satisfy his needs, and the description of the Modules of the Project and how they interact with each other

❖ Chapter 4: Implementation and Testing

This Chapter Explains the Main Projects Functions with screen shots for the project code

❖ Chapter 5: User Manual.

This Chapter describes how to operate the project and how the user will use the application to handle the input and how will get the output.

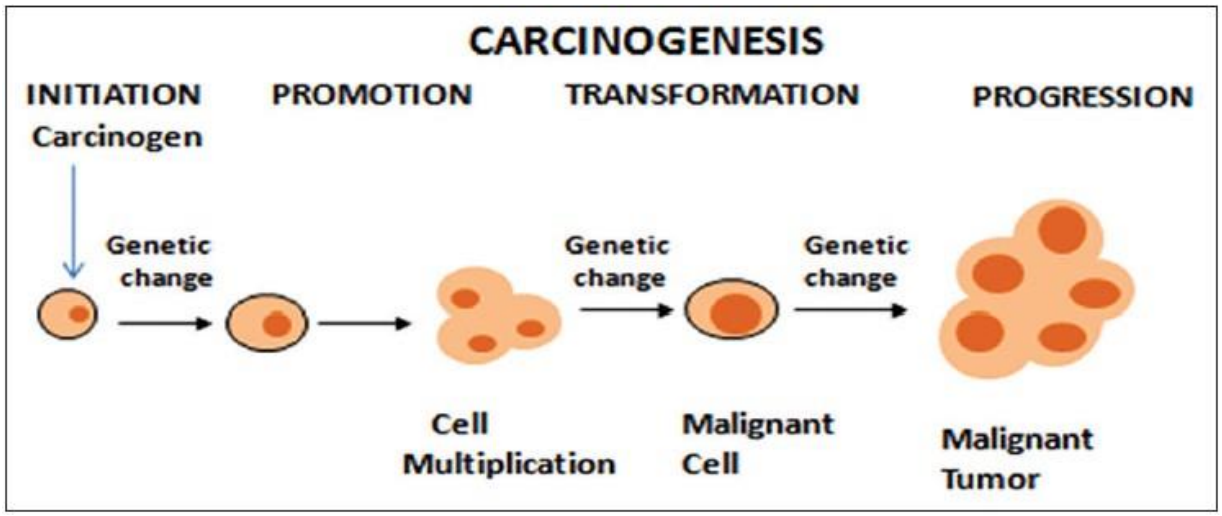
❖ Chapter 6: Conclusion and Future Work.

This Chapter finalizes the documentation with a brief summary of what was discussed in the documentation and the results achieved it also mentions our Future plans for the Project.

Chapter 2
Background

2.1 The field of the project

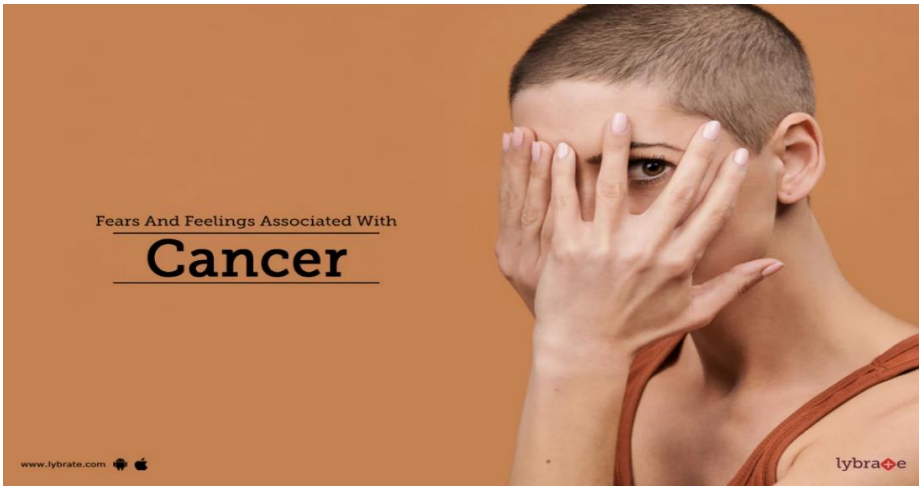
All cancers originate from a single cell that undergoes a transformation from a normally functioning somatic cell into a malignant neoplasm. In most cases, this transformation follows a stepwise process with the somatic cell first expanding into a precancer and, subsequently, becoming an advanced invasive cancer. The progression from a pre-malignant tumor to a malignant neoplasm is due to somatic mutations that can be traced, characterized, and genomically studied.fig[2.1]



Fig(2.1) carcinogenesis.

In this rotation, the student will evaluate the mutational burden, driver mutations, copy number changes, mutational signatures, and subclonal architecture of pre-malignant lesions and compare them to molecular events previously identified in advanced invasive cancers. The goal is to reveal the molecular events that are necessary for a precancer to convert into cancer. Independent previously generated

drug-screen datasets (e.g., Cancer Cell Line Encyclopedia) will be used to propose potential intervention strategies that can used to target these molecular events in order to halt this conversion and lead to cancer prevention.fig[2.1.2]



Fig(2.2)

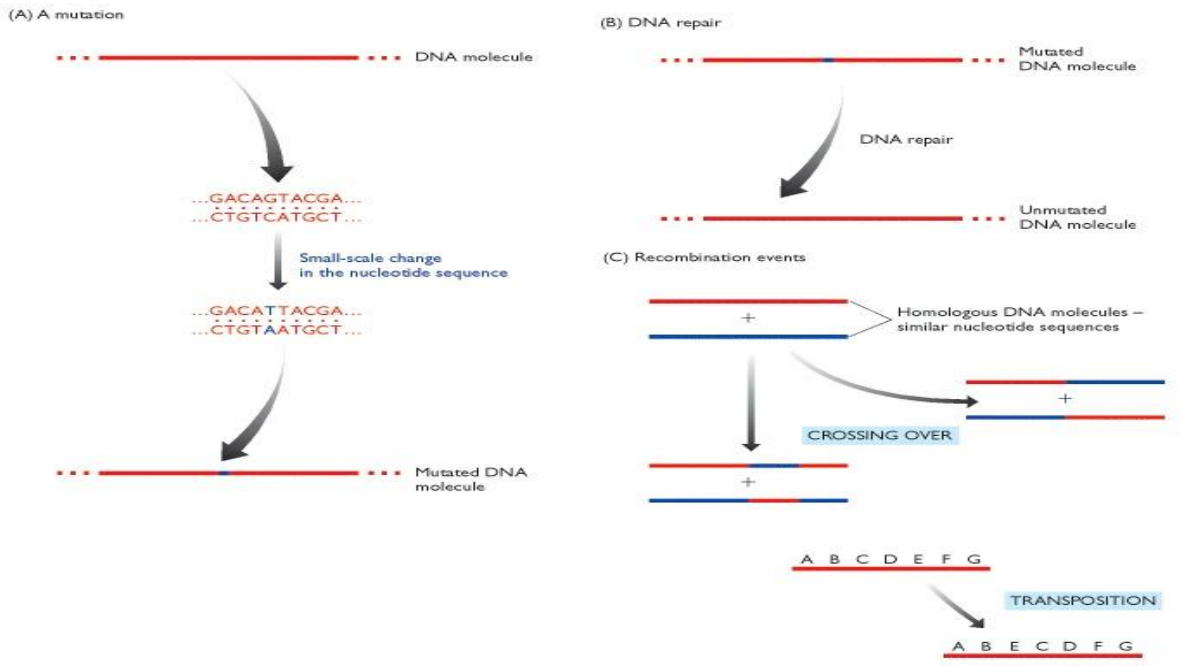
Every year, thousands of people are diagnosed with pre-cancerous conditions, news that may induce fear and panic in those receiving it. While pre-cancer that goes unchecked may ultimately become cancerous, it's not a guarantee and, in many cases, not even likely. "No one dies of pre-cancer," says Justin Chura, MD, Chief of Surgery & Director of Gynecologic Oncology and Robotic Surgery at our Philadelphia hospital. "It's a very treatable condition, if it even needs treatment at all. When patients, and even some clinicians, see the word carcinoma, they get misled into thinking they have cancer. Pre-cancer means there are cells that have grown abnormally, causing their size, shape or appearance to look different than normal cells." Whether abnormal cells become cancerous is, in many cases, an uncertainty. Some of the variables are known, others are not. So what exactly does it mean to be told you have a pre-cancerous condition? Does it increase the risk of getting cancer? Are there any prophylactic (preventative) measures that can be taken to reduce the likelihood that a cancer diagnosis is in your future? So Early

detection of oral malignant or precancerous lesion by screening individuals with high-risk factors may identify candidates who should receive treatment to prevent cancer progression and reduce patient mortality. Among the diagnostic tools, in vivo staining is advocated as a simple, inexpensive, and fairly sensitive method.

2.2 Scientific Background

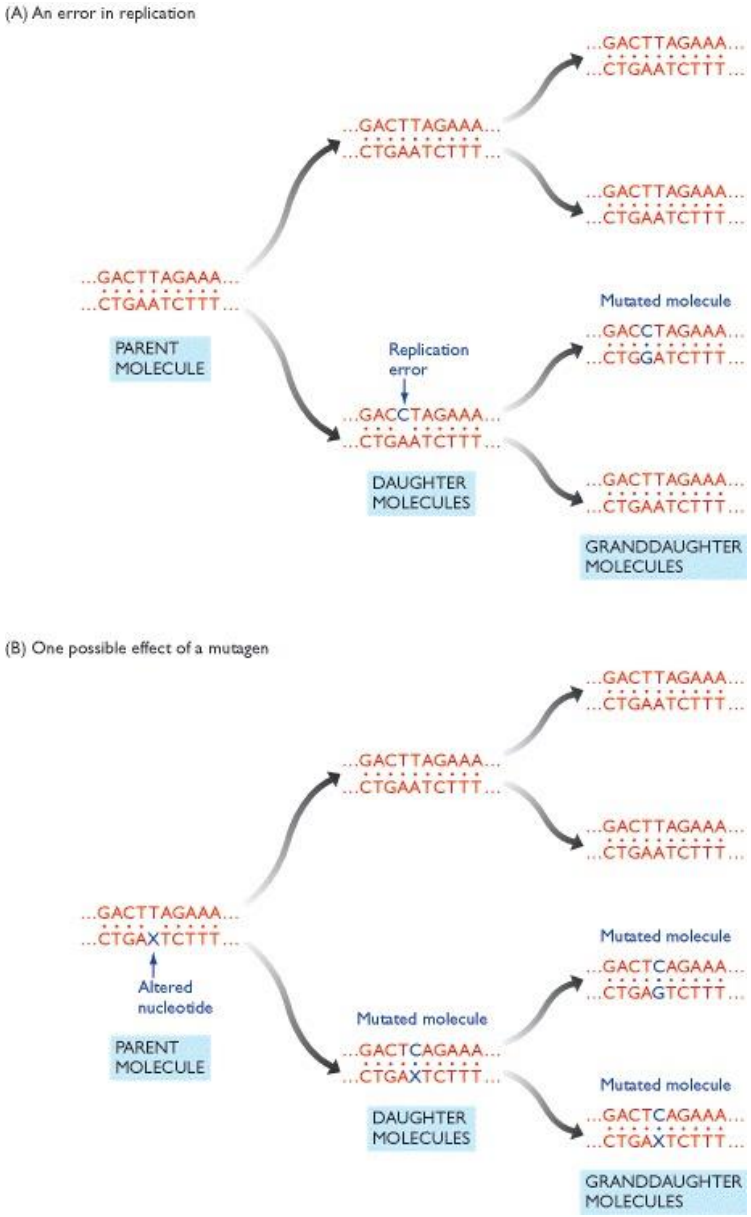
- First: What is mutation and how can mutation can cause cancer?

Based on Bioinformatics research. A mutation is a change in the nucleotide sequence of a short region of a genome fig[2.3] Many mutations are point mutations that replace one nucleotide with another; others involve insertion or deletion of one or a few nucleotides. Mutations result either from errors in DNA replication or from the damaging effects of mutagens, such as chemicals and radiation, which react with DNA and change the structures of individual nucleotides.



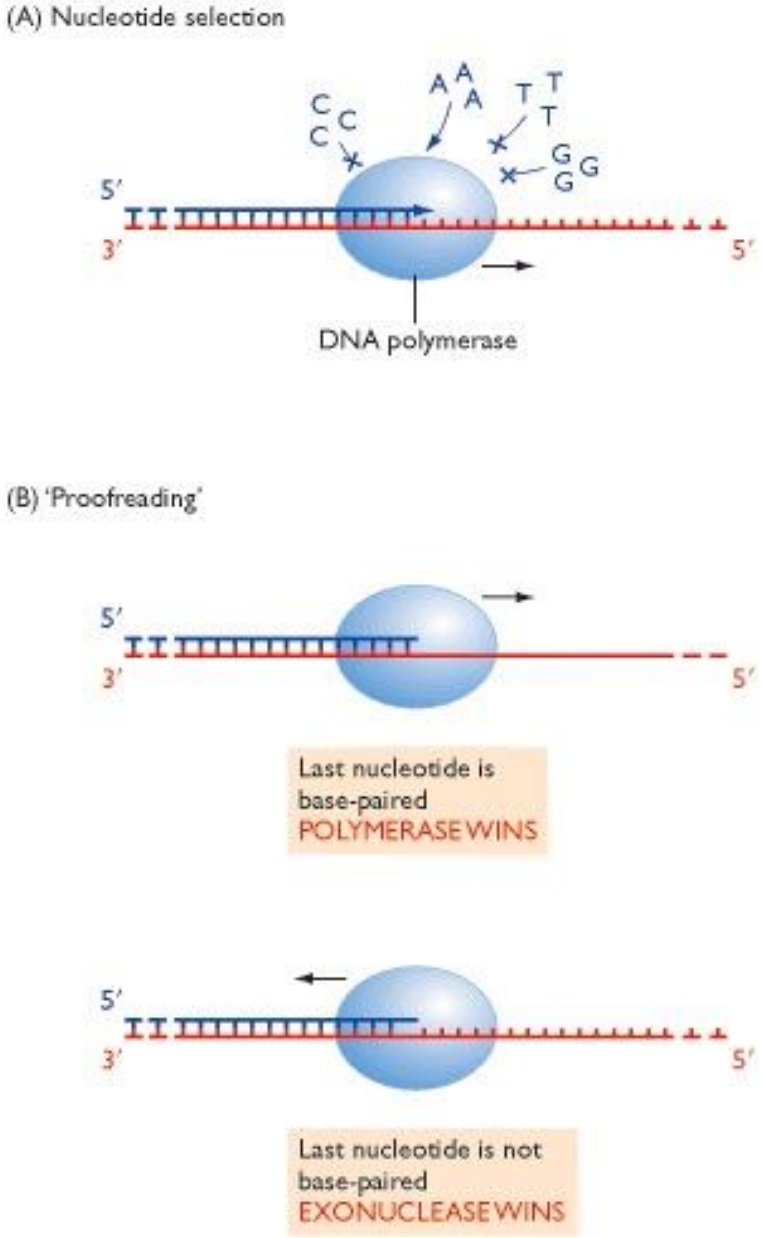
(fig2.3)Mutation, repair and recombination

All cells possess DNA-repair enzymes that attempt to minimize the number of mutations that occur. These enzymes work in two ways. Some are pre-replicative and search the DNA for nucleotides with unusual structures, these being replaced before replication occurs; others are post-replicative and check newly synthesized DNA for errors, correcting any errors that they find fig[2.4]



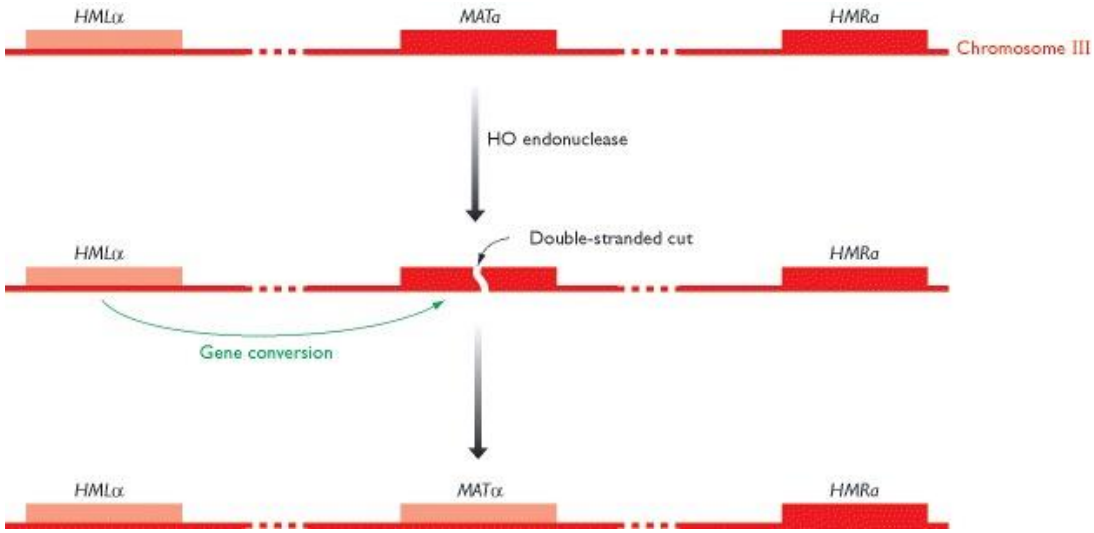
(Fig 2.4)Examples of mutations

A possible definition of mutation is therefore a deficiency in DNA repair Recombination results in a restructuring of part of a genome, for example by exchange of segments of homologous chromosomes during meiosis or by transposition of a mobile element from one position to another within a chromosome or between chromosomes fig[2.5].

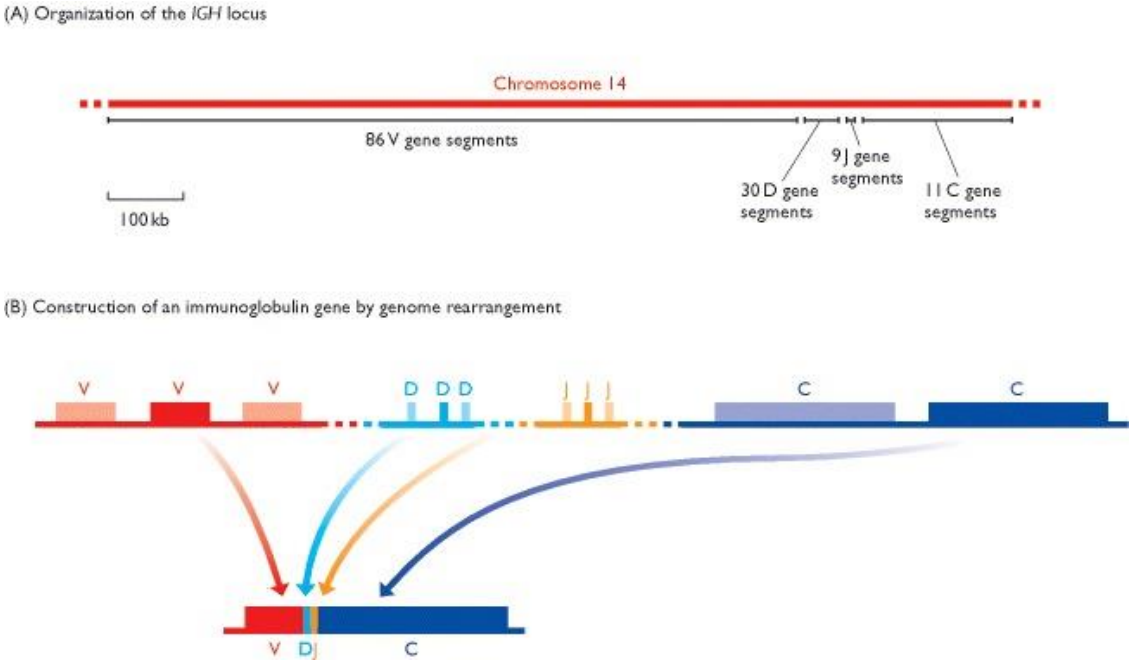


(Fig 2.5) Mechanisms for ensuring the accuracy of DNA replication

Various other events that we have studied, including mating-type switching in yeast fig[2.6]and construction of immunoglobulin genes fig[2.7] are also the results of recombination. Recombination is a cellular process which, like other cellular processes involving DNA (e.g. transcription and replication), is carried out and regulated by enzymes and other proteins.



(Fig2.6)Mating-type switching in yeast



(Fig2.7)Immunoglobulin gene segments and construction of a functional gene

Types of mutations :

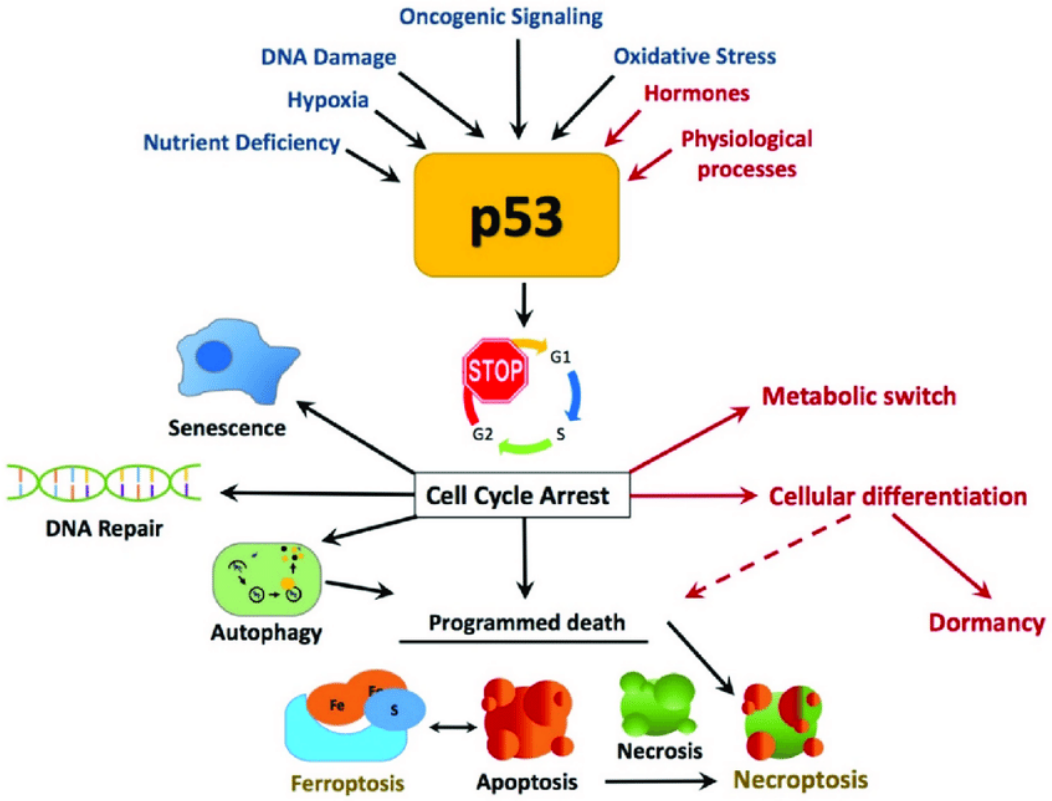
1-Some mutations are spontaneous errors in replication that evade the proofreading function of the DNA polymerases that synthesize new polynucleotides at the replication fork. These mutations are called mismatches because they are positions where the nucleotide that is inserted into the daughter polynucleotide does not match, by base-pairing, the nucleotide at the corresponding position in the template DNA. If the mismatch is retained in the daughter double helix then one of the granddaughter molecules produced during the next round of DNA replication will carry a permanent double-stranded version of the mutation.

2-Other mutations arise because a mutagen has reacted with the parent DNA, causing a structural change that affects the base-pairing capability of the altered nucleotide. Usually this alteration affects only one strand of the parent double helix, so only one of the daughter molecules carries the mutation, but two of the granddaughter molecules produced during the next round of replication will have it.

-Second : what and where is the TP53 and it's important role in repair DNA?

The TP53 gene provides instructions for making a protein called tumor protein p53 (or p53). This protein acts as a tumor suppressor, which means that it regulates cell division by keeping cells from growing and dividing (proliferating) too fast or in an uncontrolled way. The p53 protein is located in the nucleus of cells throughout the body, where it attaches (binds) directly to DNA. When the DNA in a cell becomes damaged by agents such as toxic chemicals, radiation, or ultraviolet (UV) rays from sunlight, this protein plays a critical role in determining whether the

DNA will be repaired or the damaged cell will self-destruct (undergo apoptosis). If the DNA can be repaired, p53 activates other genes to fix the damage. If the DNA cannot be repaired, this protein prevents the cell from dividing and signals it to undergo apoptosis. By stopping cells with mutated or damaged DNA from dividing, p53 helps prevent the development of tumors. Because p53 is essential for regulating DNA repair and cell division, it has been nicknamed the "guardian of the genome. "As shown in fig[2.8].



(Fig 2.8) P53 role in repairing DNA

-Third: Mutations in TP53 that cause pre-cancer and cancer stages

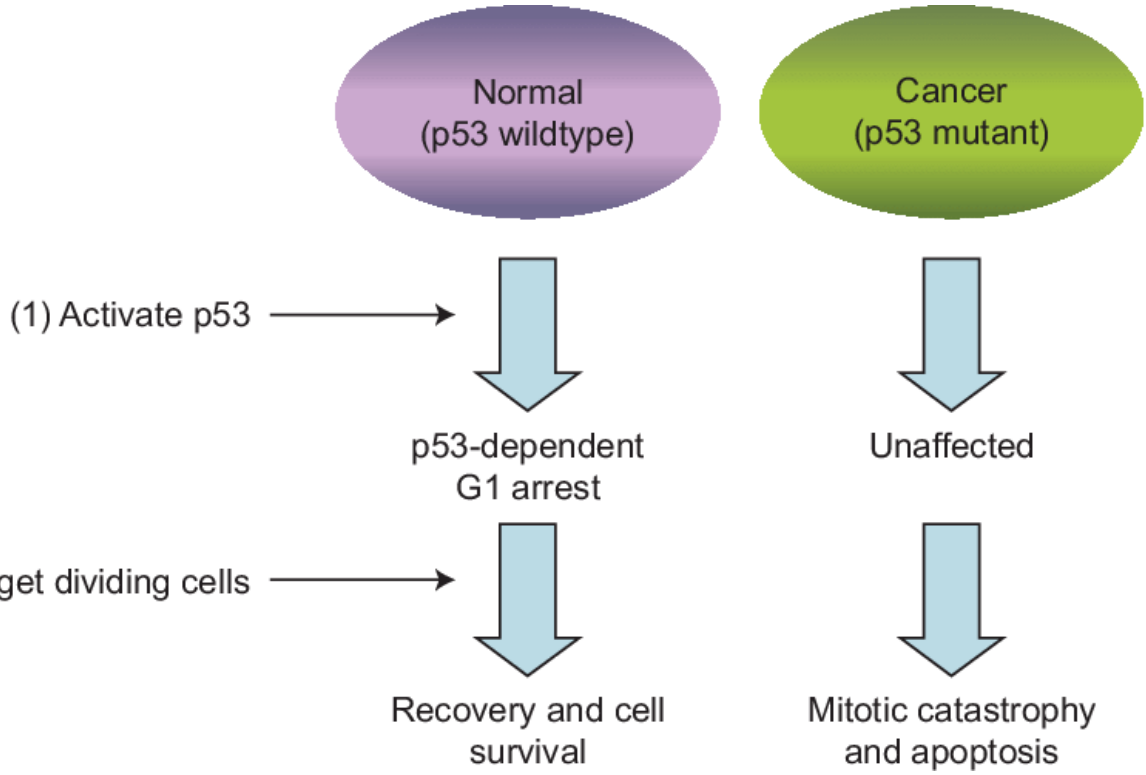
Inactivation of the p53 tumor suppressor is a frequent event in tumorigenesis. In most cases, the p53 gene is mutated, giving rise to a stable mutant protein whose accumulation is regarded as a hallmark of cancer cells. Mutant p53 proteins not only lose their tumor suppressive activities but often gain additional oncogenic

functions that endow cells with growth and survival advantages. Interestingly, mutations in the p53 gene were shown to occur at different phases of the multistep process of malignant transformation, thus contributing differentially to tumor initiation, promotion, aggressiveness, and metastasis. Here, the authors review the different studies on the involvement of p53 inactivation at various stages of digenesis and highlight the tumor specific contribution of p53 mutations at each phase of cancer and pre-cancer progression.

The notion that mutations in TP53 may occur at different stages along the process of malignant transformation raises the possibility that mutated p53 may contribute differently to various steps of this process. It is still an open question whether TP53 mutations are involved in the initiation of malignant transformation or perhaps only at more advanced stages of cancer, leading to additional growth and aggressiveness advantages. It appears, however, that the timing of the mutation during tumorigenesis is extremely variable from one cancer to another. In this review, we revisit the questions of when p53 mutations occur during malignant transformation and how these mutations affect the cancerous phenotype at different stages of tumorigenesis.

Different studies have set out to model the tumorigenesis process and describe the order of events that take place throughout this process. In the early 1990s, the Vogelstein lab used colorectal cancer (CRC) as a model system to study the sequence of genetic alterations that take place during cancer development They analyzed the different stages of CRC, starting with healthy epithelium, progressing to early, intermediate, and late adenoma and eventually carcinoma and metastasis. This analysis led them to suggest a multistep progression model. This model argues that colorectal tumorigenesis has a clonal nature and that p53 is usually

inactivated at the transition from late adenoma to carcinoma, rather than at an earlier stage. Nevertheless, the model highlights the fact that the order of the tumorigenic events may vary, whereas the combined accumulation of these changes is central. Already in that early study, several exceptions to the concept of the late timing of loss of the short arm of chromosome 17 (17p), which contains the TP53 gene, were noted: These include loss of 17p as early as in small adenomas and 17p deletions followed by other chromosomal deletions. Further evidence for the variations in the order of mutations came from additional studies. For example, while in the Vogelstein model, APC gene mutation and beta-catenin accumulation preceded the loss of chromosome 17p, another study suggested that in fact the aberrant accumulation of beta-catenin in tumors results from p53



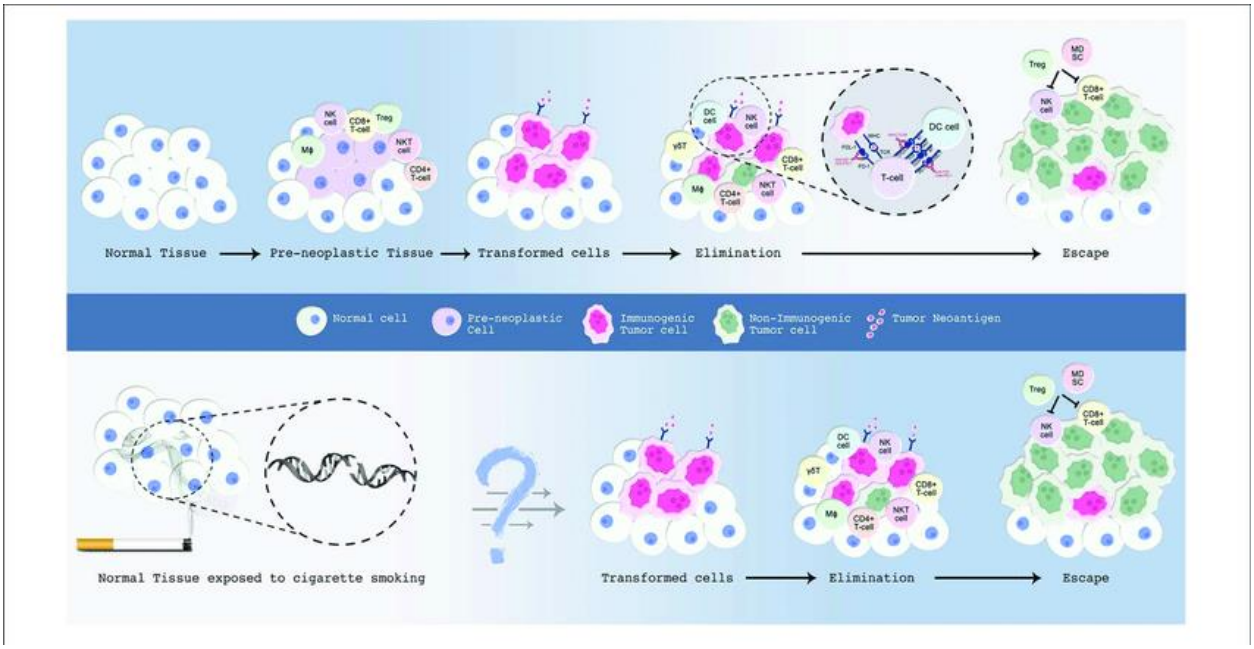
fig[2.9]

(Fig 2.9) Mutations that happens in P53 protien.

-Fourth: What is pre-cancer cells?

Pre-cancers are abnormal cells which have undergone some changes that we know are associated with an increased risk of becoming cancerous, but are not yet cancer. These changes include alterations to the inherited material (DNA) of the cells and the way those cells talk to their neighbouring cells and the immune system.

As these changes occur, the cells develop the ability to ignore normal cues that would ordinarily signal them to die (or stop replicating). This leads to an increase in the number of pre-cancer cells that may be detectable – for example, a mole on your skin made up of melanocytes (pigmented cells) or a small polyp in your colon that shows up during a colonoscopy. Additional or more powerful changes occur to switch these pre-cancers to cancer, and the chance of this happening is different for each cancer. Cancer occurs when cells become completely deaf to normal signals that constrain growth and regeneration, allowing them to start moving away from their proper location, upsetting their resident tissue and other organs around the body.fig[2.10]



(Fig2.10) Transformation of normal tissue into malignant one.

- **How much time could pre-cancer cell develop to cancer cell?**

Pre-cancers can take a long time to develop into cancer, many in fact will not progress but some can progress remarkably fast. It's this risk of cancer that clinicians and scientists are trying to prevent.

Doctors currently use a combination of the size, number, location and appearance of precancers together with molecular changes and patient history (for example exposure to known environmental risk factors such as long-term sun exposure, or a genetic susceptibility such as family history of bowel cancer) to predict the overall likelihood of cancer developing.

- **Factors causes pre-cancer:**

-Mutation in tp53 by the environment could cause pre-cancer to happen, there are quite a few factors that can cause cells to become precancerous. They vary depending upon the particular type of cells involved. In the past, researchers believed that the damage was done by carcinogens, or cancer-causing agents in the environment, that transform healthy cells into abnormal ones.

-Scientists who work in a field called epigenetics are now learning that the cells in our body are more resilient than that. A host of factors, such as carcinogens, hormones, or perhaps even stress, work together. The combination is what determines how abnormal changes in a cell may progress. One way of understanding causes is to look at the possible reasons for why damage may happen to healthy cells, leading to the genetic changes that in turn drive abnormal growth and development.

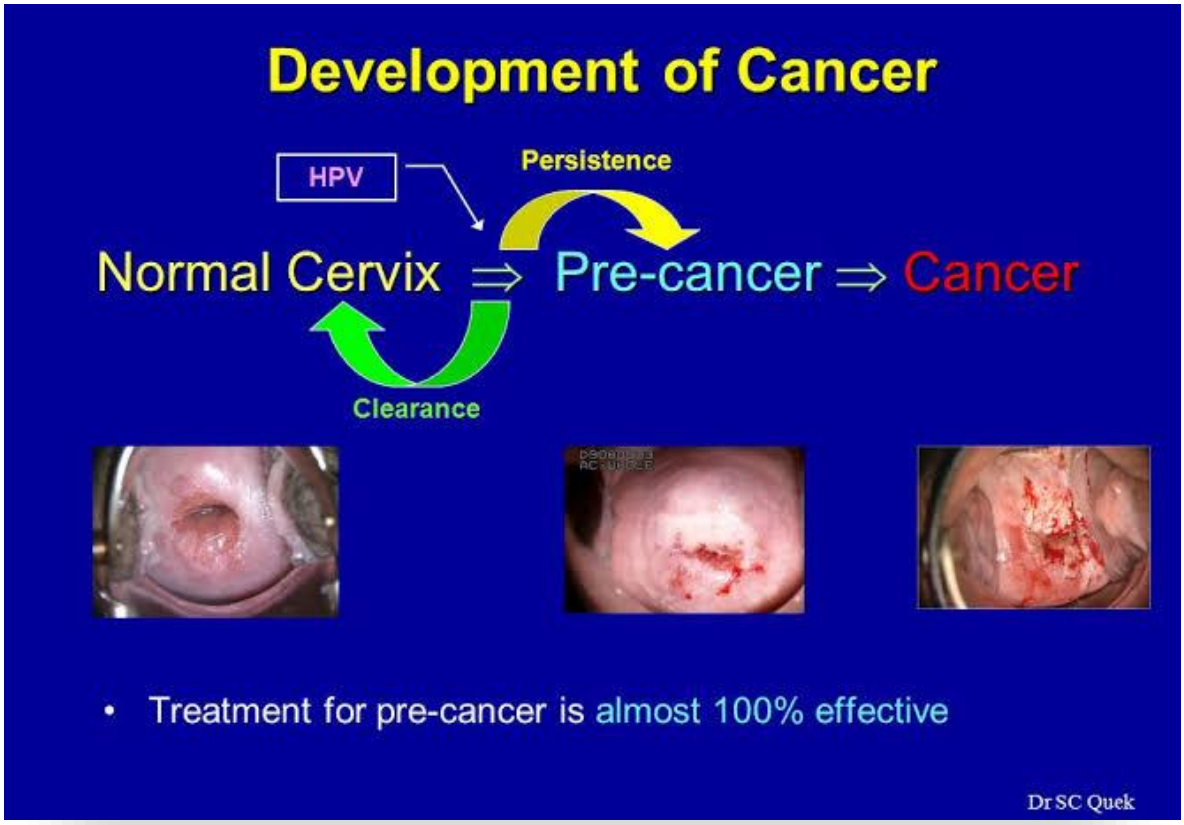
- **Main Difference Between Pre-cancer Cells, Cancer Cells And Early Stage:**

  Pre-cancer Cells: as we defined that the patient have a damaged cell and the cell have been transformed from normal cell to abnormal cell but it can be a damaged cell only. the damaged cell starts to divide slowly unlike in cancer stage and start to spread in the same position.

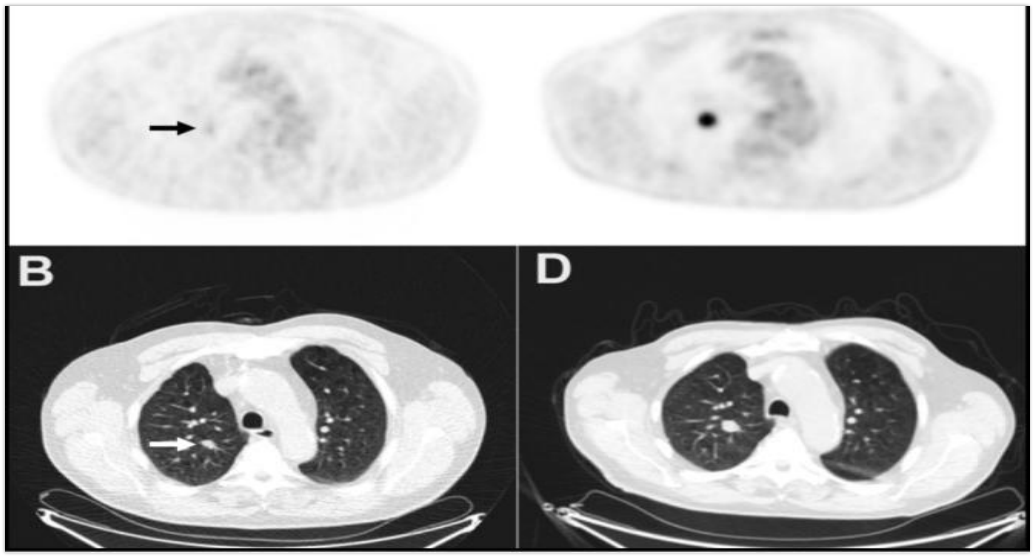
  Cancer Cells: the cell becomes damaged much more than pre-cancer one , continue to spread more and more in all positions and the cell does multiple divisions.

  Early Stage: means we already goes in the cancer stage but we try to cure the patient quickly before it spreads to much

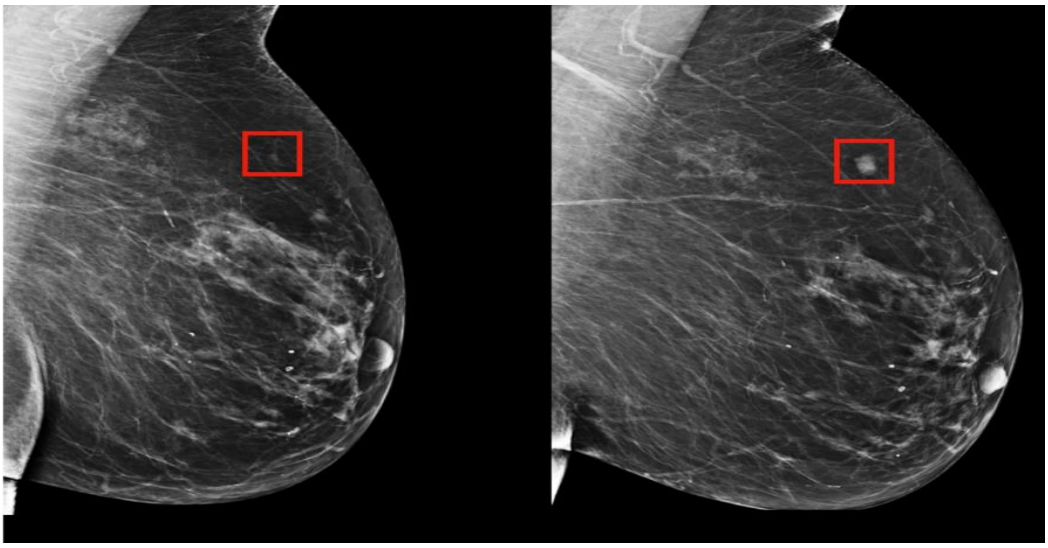
  The following figures describe how pre-cancer cells differs from cancer and early stage fig[2.11], fig[2.12], fig[2.13],



(fig2.11) difference between pre-cancer and cancer in cervix



(fig2.12) early stage



(fig2.13) early stage in breast cancer

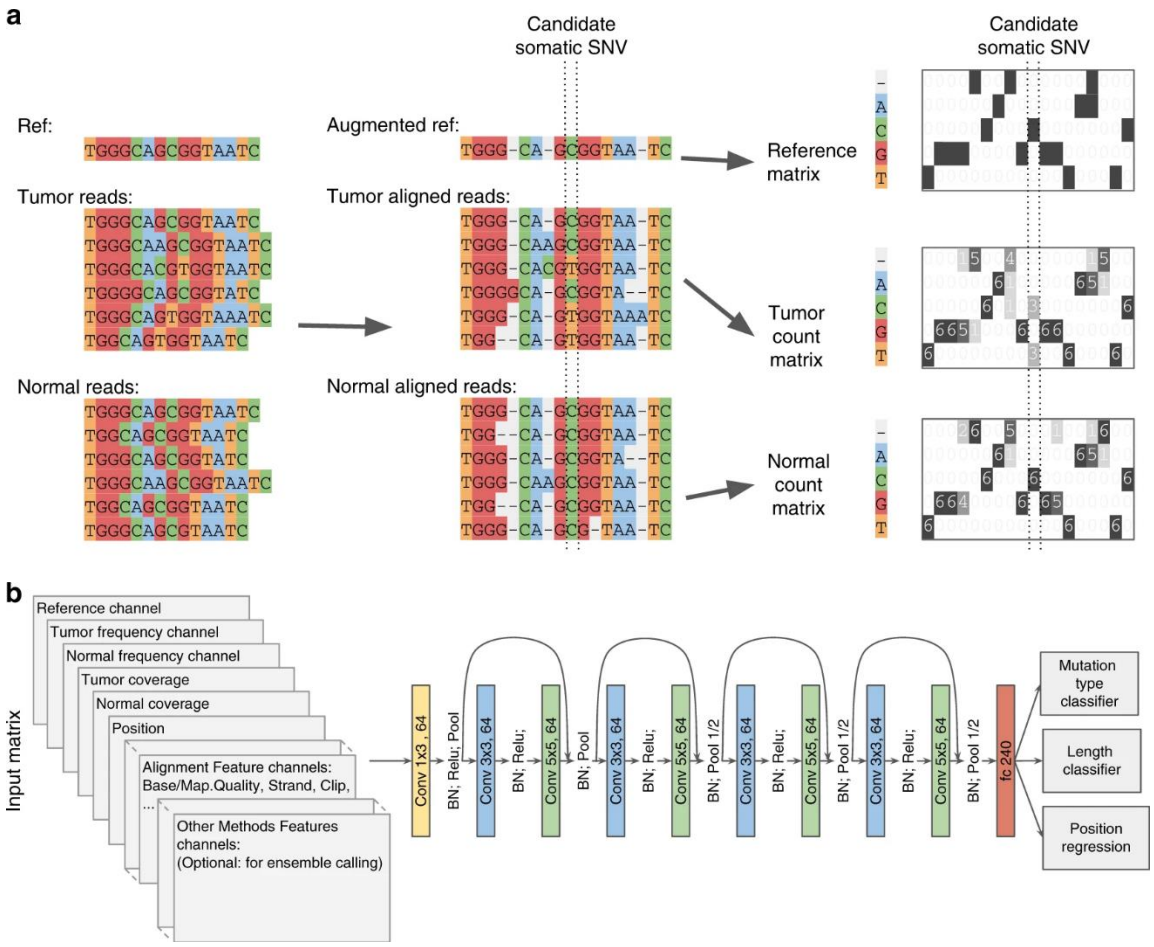
2.3 Survey of the work done & Similar systems in the field:

2.3.1 NeuSomatic:

the first CNN-based approach for somatic mutation detection that can effectively leverage signals derived from sequence alignment, as well as from other methods to accurately identify somatic mutations. Unlike other deep learning based methods that are focused on germline variants, NeuSomatic is addressing a bigger unmet need in terms of accuracy due to the complexity of tumor samples. It can effectively capture important mutation signals directly from the raw data and consistently achieve high accuracy for different sequencing technologies, sample purities, and sequencing strategies such as whole-genome vs. target enrichment.

The inputs to NeuSomatic's network are candidate somatic mutations identified by scanning the sequence alignments for the tumor sample as well as the matched normal sample fig[2.14] Somatic mutations reported by other methods can also be included in this list of candidates. For each candidate

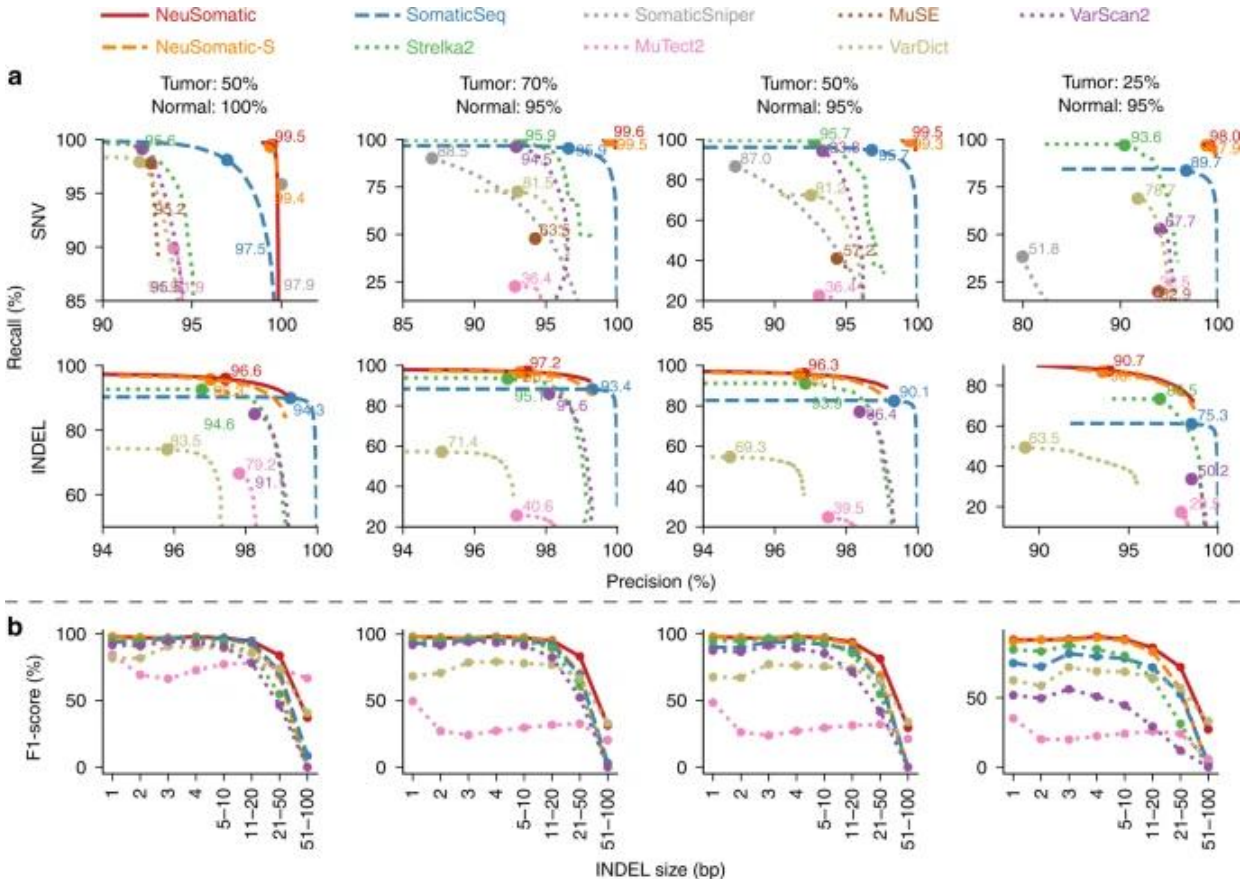
locus, we construct a 3-dimensional feature matrix M (of size k × 5 × 32), consisting of k channels each of size 5 × 32, to capture signals from the region centered around that locus. Each channel of the matrix M has five rows representing the four nucleotide bases as well as the gapped base ('−'), and 32 columns representing the alignment columns around the candidate location.



(Fig 2.14) Neusomatic overview

The first three channels, respectively, are the reference, tumor-frequency, and normal-frequency channels that summarize the reference bases around the candidate locus, as well as the frequency of different bases in that region. We augment the reference sequence around the candidate locus with gaps

corresponding to the insertions in the read in order to capture the insertions. Thus, each column of tumor and normal-frequency matrices represents the frequency of A/C/G/T/gap bases in the corresponding multiple sequence alignment (MSA) column of the tumor and normal samples, respectively. The remaining channels summarize other features, such as coverage, base quality, mapping quality, strand-bias, and clipping information for reads supporting different bases. If NeuSomatic is used in the ensemble mode, we also use additional channels for features reported by the individual somatic mutation detection methods. With this concise, yet comprehensive structured representation, NeuSomatic can use the necessary information in tumor, normal, and reference to differentiate difficult to catch somatic mutations with low AF from germline variants, as well as sequencing errors. This design also enables the use of convolutional filters in the CNN to capture contextual patterns in the sub-blocks of the matrix.



(Fig 2.15) Performance analysis of platinum of mixture of two datasets

NeuSomatic employs a novel CNN structure that predicts the type and length of a candidate somatic mutation given the feature matrix **M**. The proposed CNN consists of nine convolutional layers structured in four blocks with shortcut identity connections inspired by ResNet17 but with a different formation to adapt to the proposed input structure. We use two softmax classifiers and one regressor on the final layer. The first classifier identifies whether the candidate is a non-somatic call, SNV, insertion, or deletion. The second classifier predicts the length of the somatic mutation with four classes (0 indicating non-somatic, or lengths from 1, 2, or greater than 2), and the regressor predicts the location of the somatic mutation. Using the output of these classifiers we identify the set of somatic mutations.

If the lengths of INDELs are predicted to be larger than 2, we perform a simple post-processing step on reads overlapping that position to resolve the INDEL sequence from the read alignment CIGAR string. This has been shown to perform well for data generated by Illumina sequencers. For higher error rate sequencing data, more complex local realignment post-processing is conducted to resolve the INDEL sequence.

Results:

NeuSomatic yielded up to 99.6 and 97.2% F1-scores for SNVs and INDELs, respectively, overall and an improvement of up to 7.2% over the best method in the lowest sample purity for this dataset. For the sample with 50% tumor purity, reducing normal purity from 100 to 95% had minor impact on NeuSomatic's performance (<0.3%), whereas it caused ~3% decrease in SomaticSeq accuracy.

2.3.2 Network (QPN):

the mining technique was based on training the Quick Propagation, that is a refinement of the tradional back propagation network, by using the number of nodes (283-141-1) that used as an (input - hidden – output) layers respectively.

Results:

Following results were obtained from the learning phase for (train, test, and validation set): R-squared = (0.9987), Correlation = (0.9993), with Mean of Absolute Relative Error = (0.0057).

2.3.3 Silico molecular

lassifying approch for breast and prostate cancers by the mean of Back Propagation Network, by adopting seven datasets to assess the proposed model, five datasets from the UMD TP53 database and two sets of the (IARC TP53 database), the back propagation using hybrid pattern with 5 fold cross validation and sets of validation used for prediction and classification of breast and prostate cancers of patients using molecular mutations situated in the TP53 gene.

Results:

accuracy = (98 , 96.7), specificity = (96.6 , 97.3) , and sensitivity (Recall) = (97 , 96.5).

2.3.4 Back Propagation Neural Network (BPNN) & ReliefK

ReliefK output favorable results for feature selection. Back Propagation Neural Network (BPNN) also used with this method to predict and classify cancer based on the mutations that were listed in the somatic and germline mutations of IARC TP53 database.

Results:

Matthew correlation coefficient (MCC) for both BPNN & ReliefK: 1 and 0.88 for IARC TP53 somatic and germline mutations, respectively.

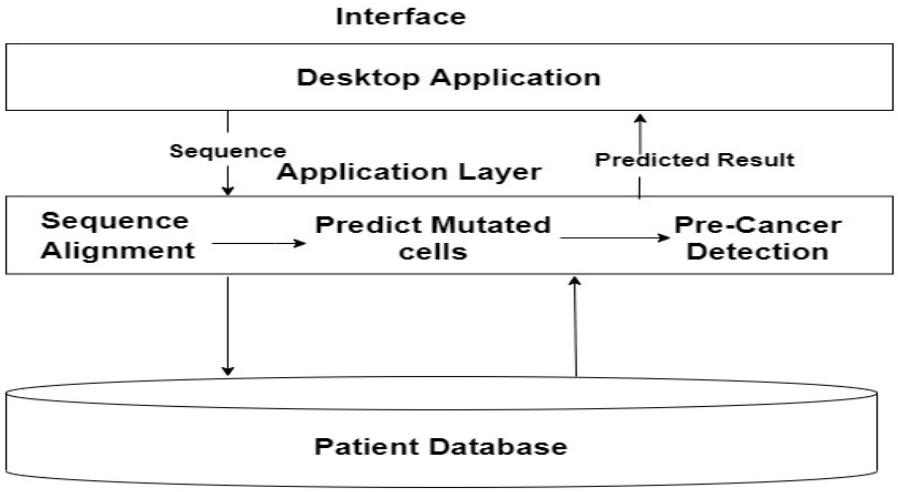
Chapter 3
Analysis and Design

3.1 System Overview

3.1.1 System Architecture

- **Interface Layer:** Desktop Application
- **Application Layer Consists of three functions:**
  - **Sequence Alignment:** Making alignment between mutated sequence that causes pre-cancer cells to be activated and normal one to know which cell have been mutated.
  - **Predicted Mutated Cells:** Show the region of mutated cells that causes pre-cancer cells.
  - **Pre-Cancer Detection:** Identify the transformation of cells into abnormal cells that's neither cancer cells nor normal cells but it's the stage before cancer cells to be developed.

- **Patient Database:** Including all patients that're normal ,have cancer and pre-cancer disease.



(Fig 3.1)System Architecture

3.1.2 Functional Requirements

-The application shall be able to make sequence alignment and detect the region of mutation.

- The application shall be able to Classify the mutated cells.

- The application shall be able to Identify Whether this mutated cells are pre-cancer cells nor cancer.

- The application should produce the results of patients.

3.1.3 Nonfunctional Requirements

- **Performance & Scalibilty:**

The system response time and how fast the system return result, How much will this performance change with higher workloads.

- **Reliability:**

Pre Cancer Detection system must have at least 90% reliability per month.

- **Security:**

the system and its data protected against attacks by stealing records of patients and doctors information from the device.

- **Usability:**

The system is very easy to use with any type of user even without usage instructions.

- **Availability**

the system is accessible for a user at any point in time.

3.1.4 System Users

A. Intended Users:

- The system was created specifically for an organization specialized in cancer researches and early detection of cancer(Pre-Cancer cells Stage).

The first group is a group of researchers specialized in the field of cancer, and the system is used through them by adding new samples of DNA from suspected patients or those who have cancer, in order to follow the conditions of the cell and the changes that occur to the genetic content of mutations during its transformation into a pre-cancer cell.

The second group is a group of doctors who're specialized in telling the patient about his condition and who will start the treatment journey earlier.

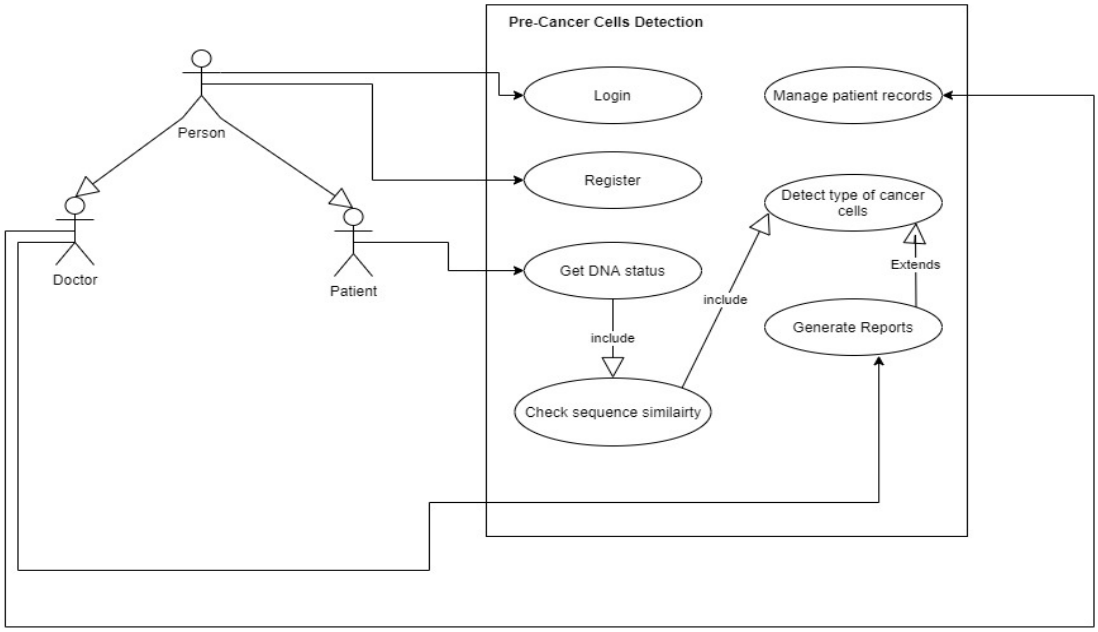
B. User Characteristics

The User must has:

- Extensive experience in the use of computers and its various programs.
- Extensive medical and research experience in the field of Pre cancer cells and an elaborate biological background in the types of mutations and how to deal with each of them.

3.2 System Analysis & Design

3.2.1 Use Case Diagram



(Fig 3.2) use case diagram

- **Login & Register**

  Description: The user register to the system if he/she doesn't have any previous account and in case of login he/she already has an account and only get into the system.

- **Check Sequence Similarity**

  Description: The system sees the alignment between abnormal cells that's mutated and normal one to check where's the mutation has been detected.

- **Get DNA Status**

  Description: The system gives the user the result of his/her check after alignment.

- **Detect type of cancer cells**

  Description: Identifying the type of mutated cells if it's a pre-cancer cells or converted into cancer cells.

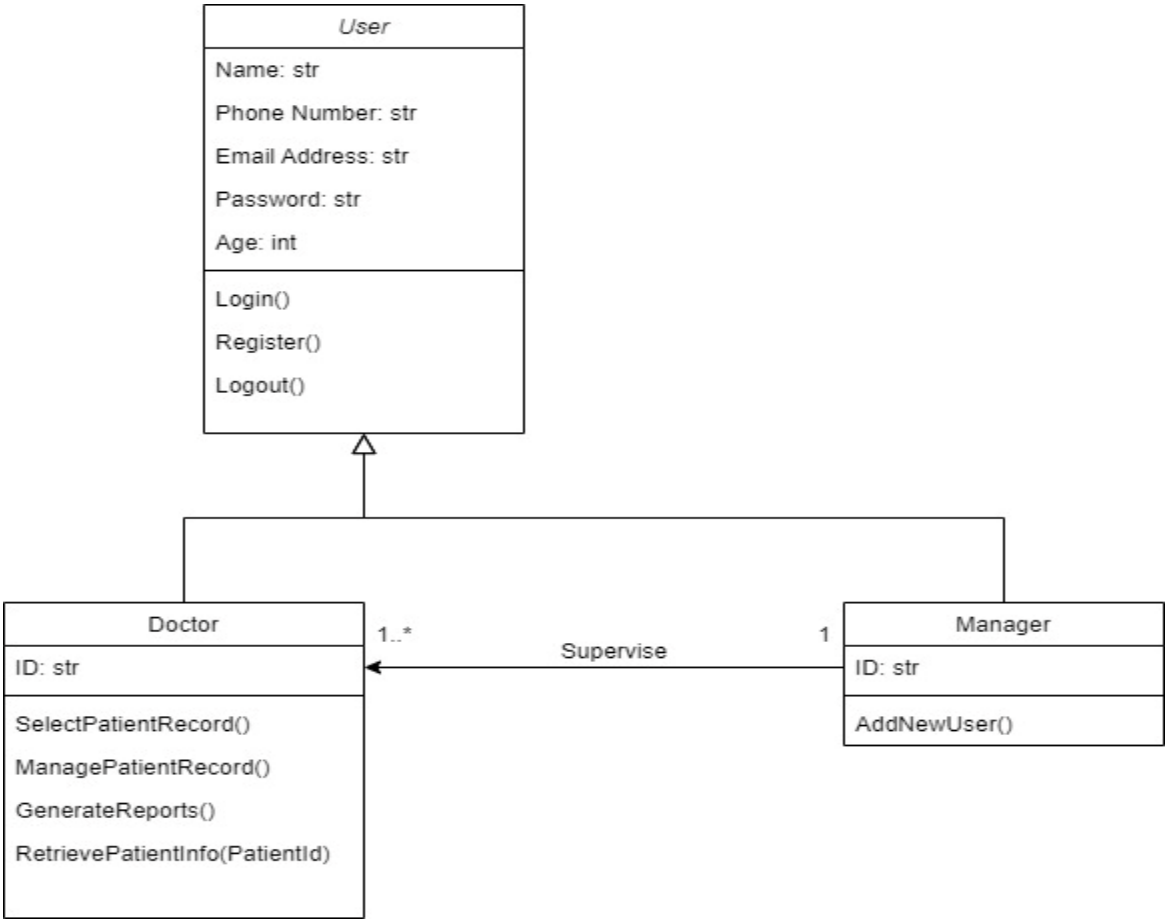
- **Manage patient records**

  Description: The doctor in the system check all records of patients to see who's urgent (have pre-cancer cells converted into cancer one).

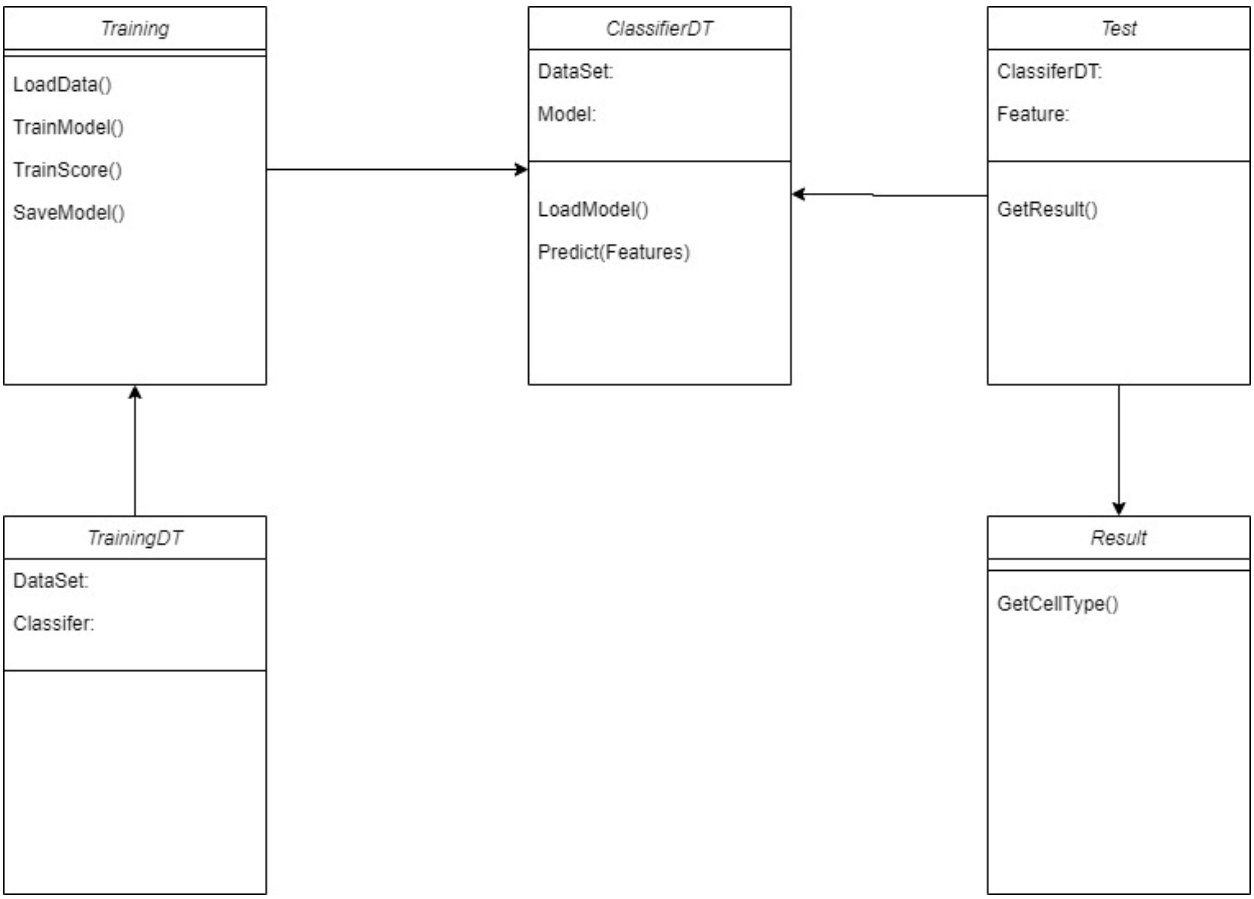
- **Generate Reports**

  Description: The doctor in the system gives the patient the report that includes his condition.

3.2.2 Class Diagram

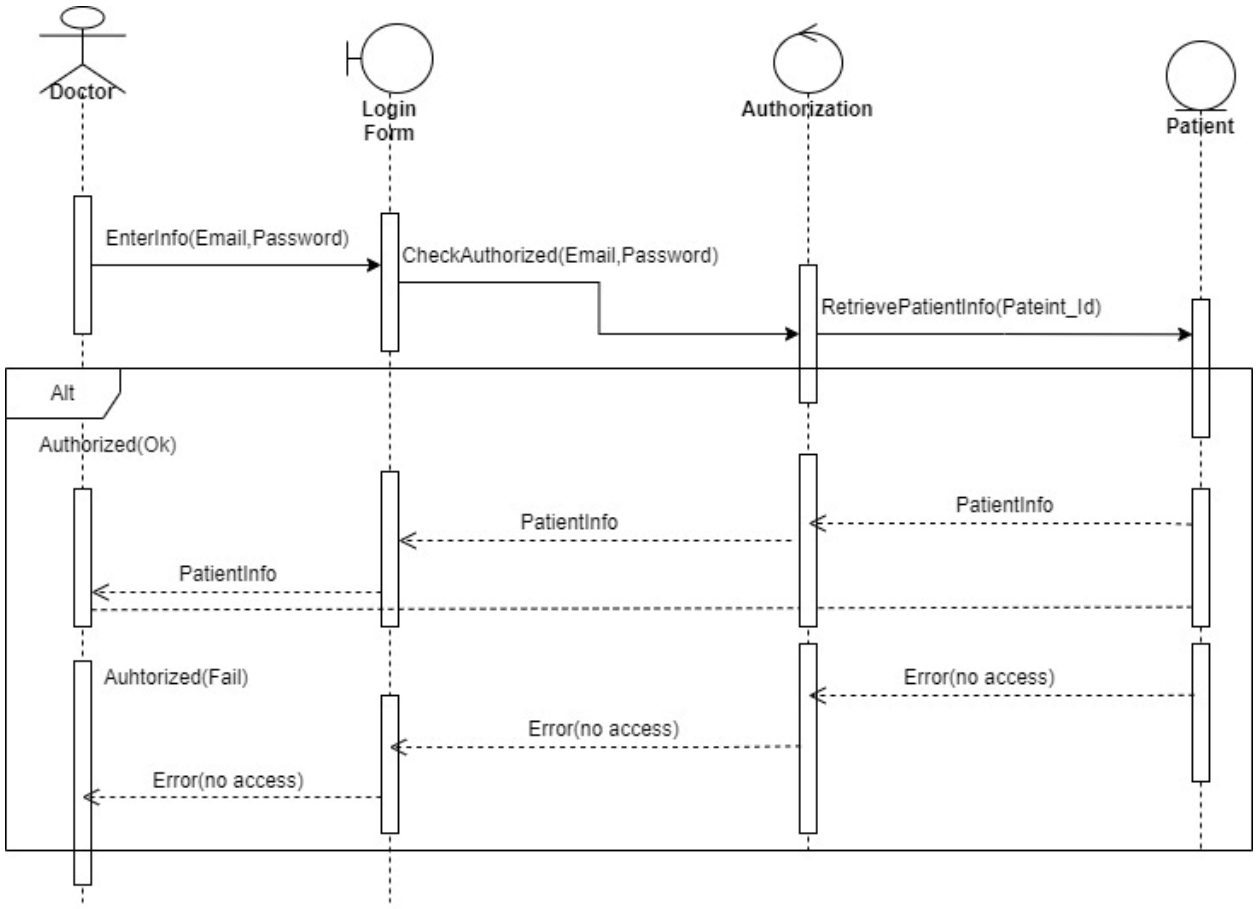


(Fig 3.3) Class diagram

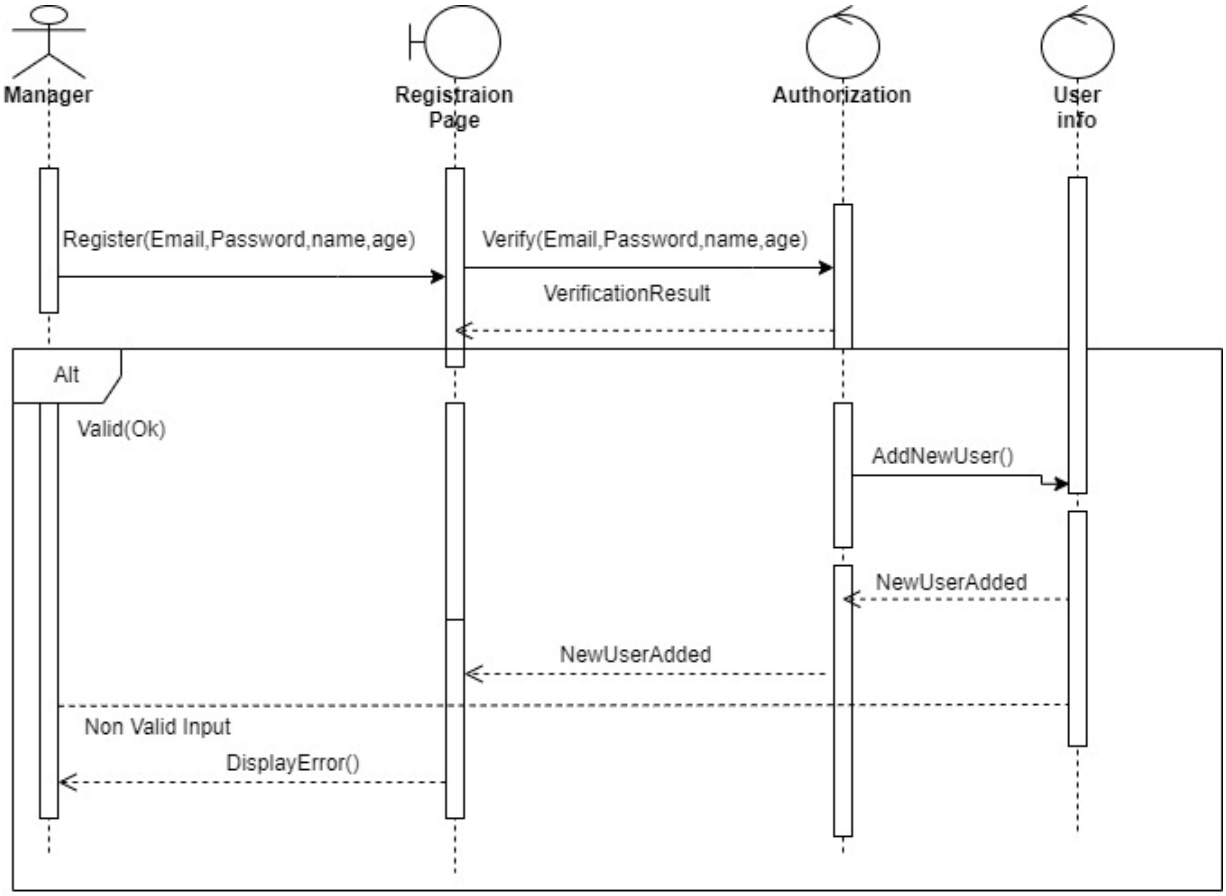


(Fig 3.4) Class diagram

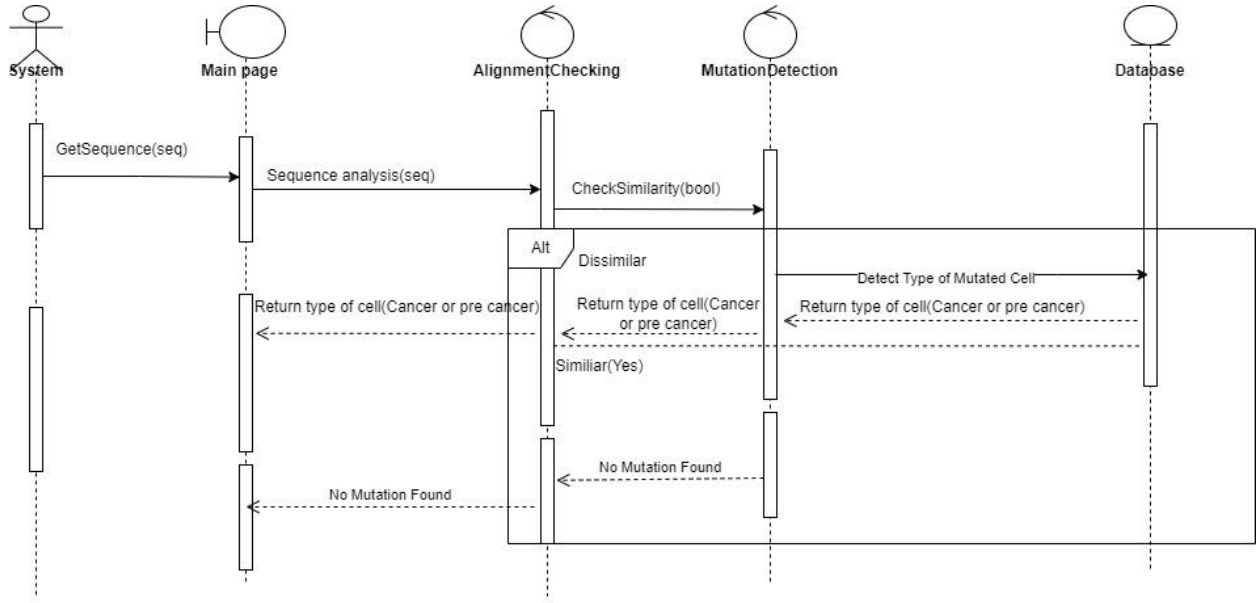
3.2.3 Sequence Diagrams



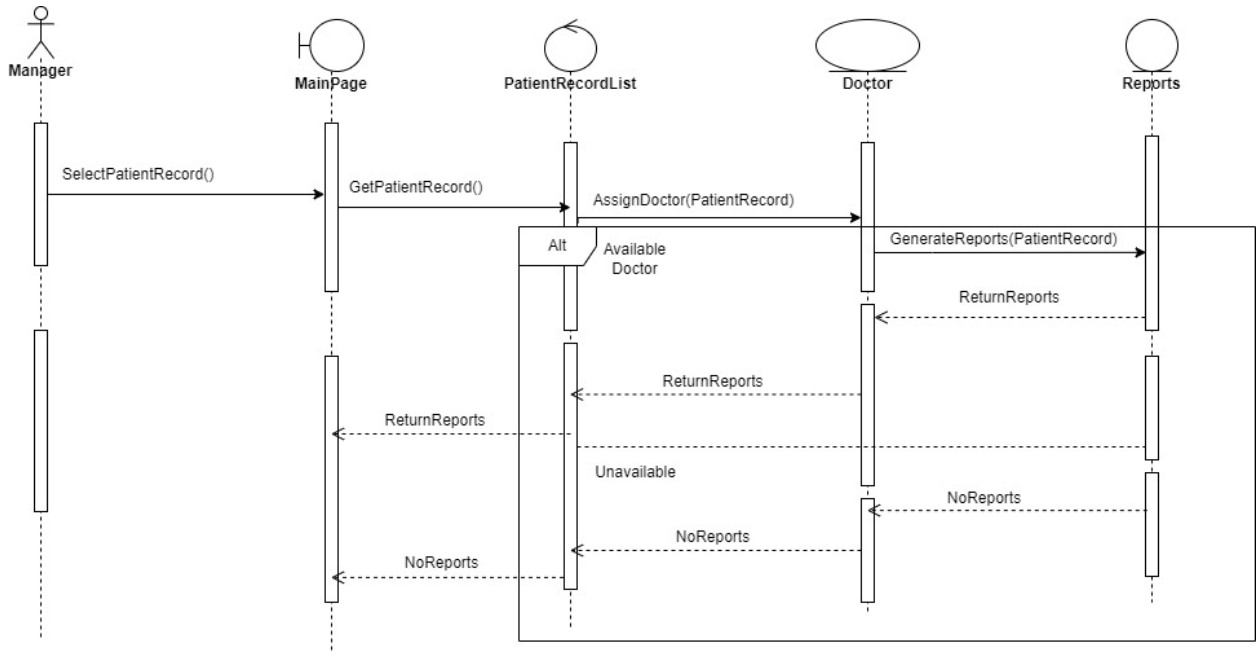
(Fig 3.5) Sequence Diagram for login



(Fig 3.6) Sequence diagram for register



(Fig 3.7) Sequence Diagram for getting DNA status, sequence similarity and detection of mutation



(Fig 3.8) Sequence diagram for manage & generate reports.

Chapter 4
Implementation

- Database: UMD (Universal Mutation Database)

  Marseille is also the home of the UMD databases. These tools are dedicated to the collection of mutations in human genes associated with genetic diseases. Most of these locus specific databases are freely accessible but some can only be accessed by a password.

  The human genome contains about 80,000 genes and presently only 3,000 are known to be implicated in genetic diseases. In the near future, the entire sequence of the human genome (Human Genome Project) will be available and the development of new methods for point mutation detection will lead to a huge increase in the identification of genes and their mutations associated with genetic diseases as well as cancers.

  The collection of these mutations will be critical for researchers and clinicians to establish genotype/phenotype correlations. Other fields such as molecular epidemiology will also be developed using these new data. Consequently, the future lies not in simple repositories of locus-specific mutations but in dynamic databases linked to various computerized tools for their analysis and that can be directly queried on-line. To meet this goal, There is a generic software designed specially to construct dataset called UMD (Universal Mutation Database).

  The UMD central tool is designed to query multiple Locus Specific Databases (LSDB) developed with the Universal Mutation Database software (UMD). Because each LSDB contain molecular and clinical data from patients with a mutation in a specific gene, they are very useful for

clinicians (phenotype-genotype correlations...), geneticists (distribution and frequency of mutations...) and research biologists (structural domains, molecular epidemiology...). Nevertheless, in some situations, only the combination of data from various databases may allow a comprehensive analysis.

The *TP53* Database compiles *TP53* variant data that have been reported in the published literature since 1989 or are available in other public databases. Database releases are identified by a number. The current is R20 (July 2019) fig[4.1]

The *TP53* Database compiles various types of data and information from the literature

And generalist

databases on human *TP53* gene variants related to cancer.

Data are organized in datasets that can be searched and analyzed with graph tools or

Fully downloaded to perform custom analyses.

release compiles data on around 29900 somatic variants, 9200 variants reported in SNP

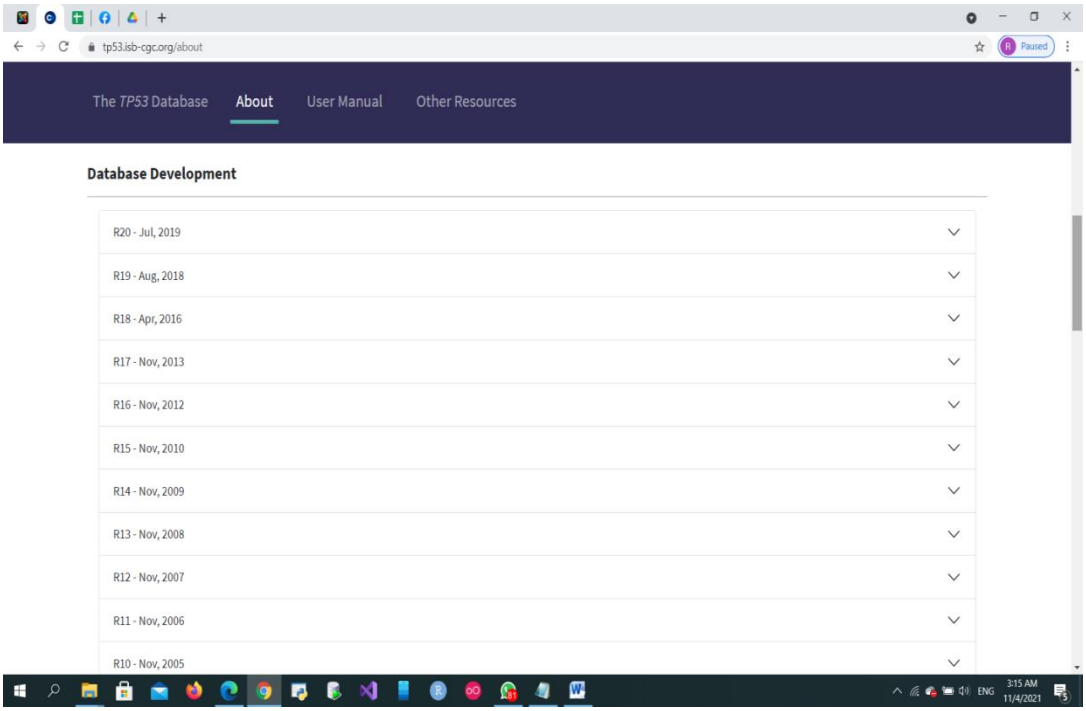
databases**,** 1530 cancer families/individual carriers of a germline variant, 2700 cell-lines,

900 experimentally induced variants, and functional data on over 9000 mutant proteins.

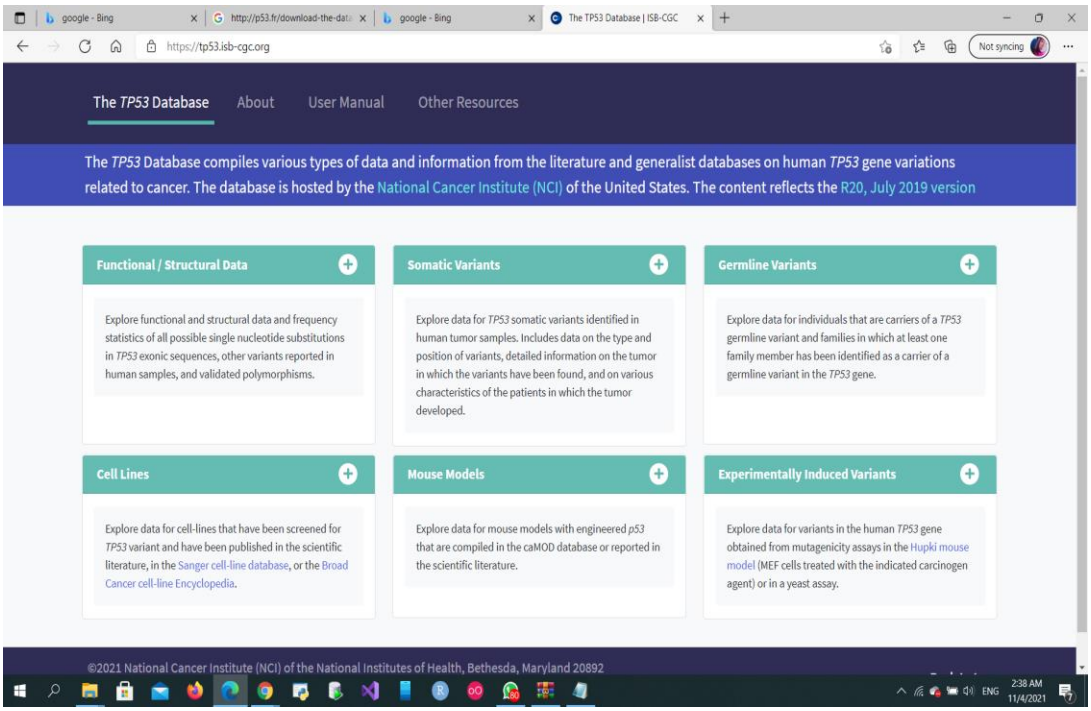
Variant descriptions are provided on both hg19 and hg38 genome builds fig[4.2].

The original IARC *TP53* database was maintained by WHO's International Agency for Research on Cancer, Lyon, France, and that responsibility has now been fully transferred to NCI. The data contained herein may be freely

used, downloaded and reproduced, but are not for sale or for use in conjunction with commercial or promotional purposes, and any use shall be subject to an appropriate acknowledgement of the source.



(Fig4.1) All database development from R1:R20

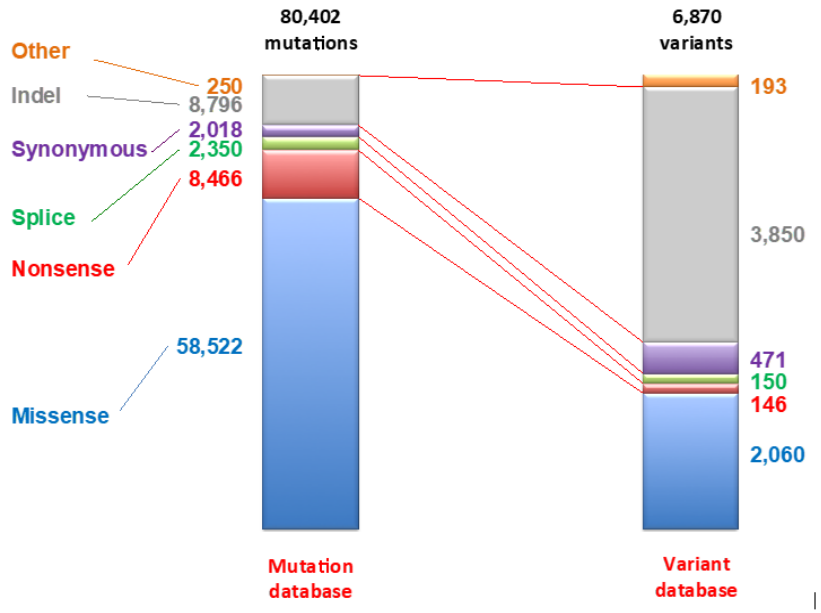


(Fig4.2) Different database categories

The UMD database comes in two files, the variant and the mutation database fig[4.3]

The mutation database includes all patients carrying a TP53 mutation. Therefore, different patients expressing the same TP53 variant are included in this database.

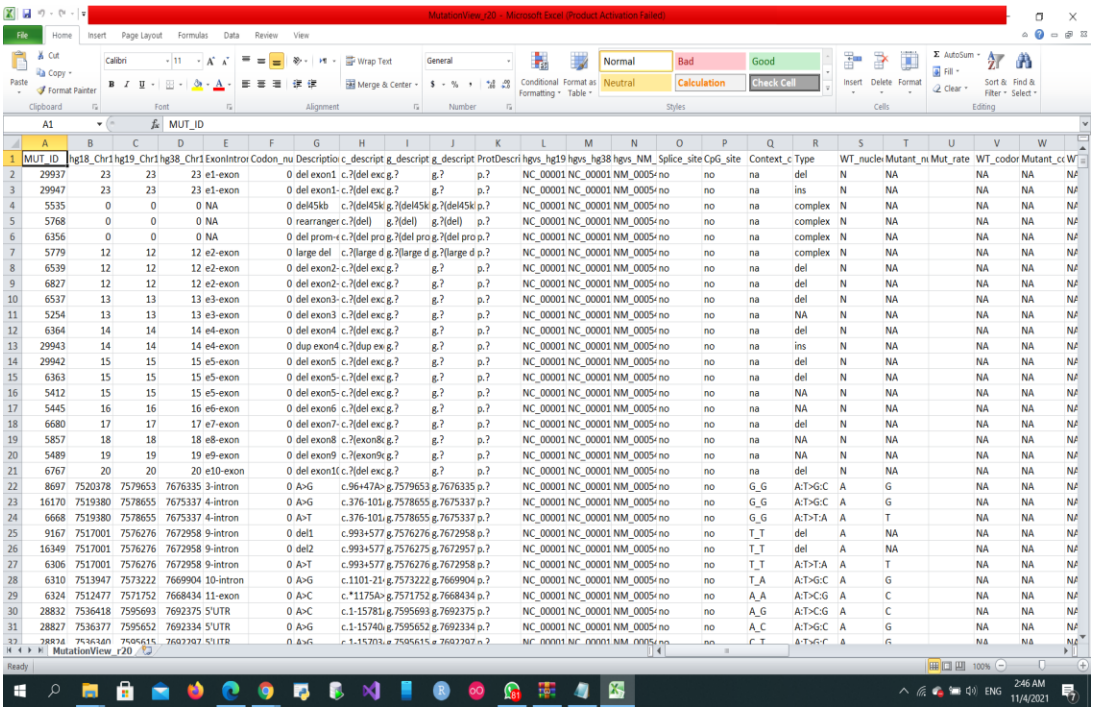
The variant database includes each single TP53 variants found in the cases database.



Relationship between the variant and the mutation database (numbers can be slighly different in the files due to an update of the database)

(Fig4.3)Relationship between variant and mutation databases

In our project (Detection of pre-Cancer cell) we use the mutation database only, which construct of 80457 row (data record or gene) and 132 row (Feature) in CSV file format. fig[4.4]

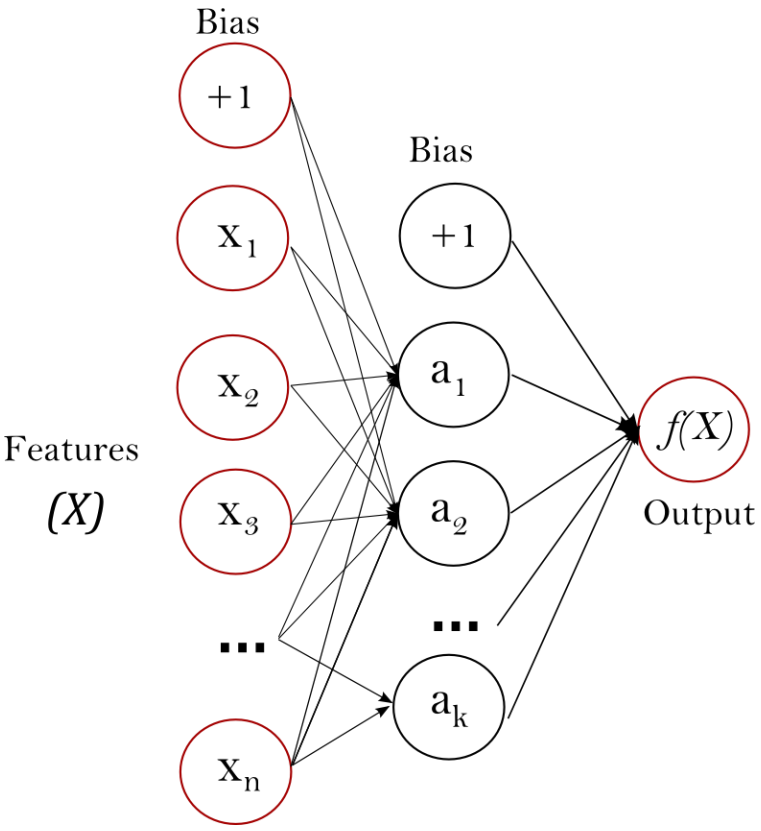


MUT_ID	hg18_Chr	hg19_Chr	hg38_Chr	Exon	Intron	Codon_nu	Description	c_descript	g_descript	g_descript	ProtDescri	hgvs_hg19	hgvs_hg38	hgvs_NM_	Splice_site	CpG_site	Context_c	Type	WT_nucle	Mutant_n	Mut_rate	WT_codor	Mutant_c
29937	23	23	23	e1	exon	0	del exon1	c.?(del exc	g.?	g.?	p.?	NC_00001	NC_00001	NM_0005	no	no	na	del	N	NA		NA	NA
29947	23	23	23	e1	exon	0	del exon1-	c.?(del exc	g.?	g.?	p.?	NC_00001	NC_00001	NM_0005	no	no	na	ins	N	NA		NA	NA
5535	0	0	0	NA		0	del45kb	c.?(del45kb	g.?(del45kb	g.?(del45kb	p.?	NC_00001	NC_00001	NM_0005	no	no	na	complex	N	NA		NA	NA
5768	0	0	0	NA		0	rearranger	c.?(del)	g.?(del)	g.?(del)	p.?	NC_00001	NC_00001	NM_0005	no	no	na	complex	N	NA		NA	NA
6356	0	0	0	NA		0	del prom-	c.?(del pro	g.?(del pro	g.?(del pro	p.?	NC_00001	NC_00001	NM_0005	no	no	na	complex	N	NA		NA	NA
5779	12	12	12	e2	exon	0	large del	c.?(large d	g.?(large d	g.?(large d	p.?	NC_00001	NC_00001	NM_0005	no	no	na	complex	N	NA		NA	NA
6539	12	12	12	e2	exon	0	del exon2-	c.?(del exc	g.?	g.?	p.?	NC_00001	NC_00001	NM_0005	no	no	na	del	N	NA		NA	NA
6827	12	12	12	e2	exon	0	del exon2-	c.?(del exc	g.?	g.?	p.?	NC_00001	NC_00001	NM_0005	no	no	na	del	N	NA		NA	NA
6537	13	13	13	e3	exon	0	del exon3-	c.?(del exc	g.?	g.?	p.?	NC_00001	NC_00001	NM_0005	no	no	na	del	N	NA		NA	NA
5254	13	13	13	e3	exon	0	del exon3	c.?(del exc	g.?	g.?	p.?	NC_00001	NC_00001	NM_0005	no	no	na	NA	N	NA		NA	NA
6364	14	14	14	e4	exon	0	del exon4	c.?(del exc	g.?	g.?	p.?	NC_00001	NC_00001	NM_0005	no	no	na	del	N	NA		NA	NA
29943	14	14	14	e4	exon	0	dup exon4	c.?(dup exc	g.?	g.?	p.?	NC_00001	NC_00001	NM_0005	no	no	na	ins	N	NA		NA	NA
29942	15	15	15	e5	exon	0	del exon5	c.?(del exc	g.?	g.?	p.?	NC_00001	NC_00001	NM_0005	no	no	na	del	N	NA		NA	NA
6363	15	15	15	e5	exon	0	del exon5-	c.?(del exc	g.?	g.?	p.?	NC_00001	NC_00001	NM_0005	no	no	na	del	N	NA		NA	NA
5412	15	15	15	e5	exon	0	del exon5-	c.?(del exc	g.?	g.?	p.?	NC_00001	NC_00001	NM_0005	no	no	na	NA	N	NA		NA	NA
5445	16	16	16	e6	exon	0	del exon6	c.?(del exc	g.?	g.?	p.?	NC_00001	NC_00001	NM_0005	no	no	na	NA	N	NA		NA	NA
6680	17	17	17	e7	exon	0	del exon7-	c.?(del exc	g.?	g.?	p.?	NC_00001	NC_00001	NM_0005	no	no	na	del	N	NA		NA	NA
5857	18	18	18	e8	exon	0	del exon8	c.?(exon8(	g.?	g.?	p.?	NC_00001	NC_00001	NM_0005	no	no	na	NA	N	NA		NA	NA
5489	19	19	19	e9	exon	0	del exon9	c.?(exon9(	g.?	g.?	p.?	NC_00001	NC_00001	NM_0005	no	no	na	NA	N	NA		NA	NA
6767	20	20	20	e10	exon	0	del exon1(	c.?(del exc	g.?	g.?	p.?	NC_00001	NC_00001	NM_0005	no	no	na	del	N	NA		NA	NA
8697	7520378	7579653	7676335	3	intron	0	A>G	c.96+47A>	g.7579653	g.7676335	p.?	NC_00001	NC_00001	NM_0005	no	no	G_G	A:T>G:C	A	G		NA	NA
16170	7519380	7578655	7675337	4	intron	0	A>G	c.376-101	g.7578655	g.7675337	p.?	NC_00001	NC_00001	NM_0005	no	no	G_G	A:T>G:C	A	G		NA	NA
6668	7519380	7578655	7675337	4	intron	0	A>T	c.376-101	g.7578655	g.7675337	p.?	NC_00001	NC_00001	NM_0005	no	no	G_G	A:T>T:A	A	T		NA	NA
9167	7517001	7576276	7672958	9	intron	0	del1	c.993+577	g.7576276	g.7672958	p.?	NC_00001	NC_00001	NM_0005	no	no	T_T	del	A	NA		NA	NA
16349	7517001	7576276	7672958	9	intron	0	del2	c.993+577	g.7576275	g.7672957	p.?	NC_00001	NC_00001	NM_0005	no	no	T_T	del	A	NA		NA	NA
6306	7517001	7576276	7672958	9	intron	0	A>T	c.993+577	g.7576276	g.7672958	p.?	NC_00001	NC_00001	NM_0005	no	no	T_T	A:T>T:A	A	T		NA	NA
6310	7513947	7573222	7669904	10	intron	0	A>G	c.1101-21	g.7573222	g.7669904	p.?	NC_00001	NC_00001	NM_0005	no	no	T_A	A:T>G:C	A	G		NA	NA
6324	7512477	7571752	7668434	11	exon	0	A>C	c.*1175A>	g.7571752	g.7668434	p.?	NC_00001	NC_00001	NM_0005	no	no	A_A	A:T>C:G	A	C		NA	NA
28832	7536418	7595693	7692375	5'UTR		0	A>C	c.1-15781	g.7595693	g.7692375	p.?	NC_00001	NC_00001	NM_0005	no	no	A_G	A:T>C:G	A	C		NA	NA
28827	7536377	7595652	7692334	5'UTR		0	A>G	c.1-15740	g.7595652	g.7692334	p.?	NC_00001	NC_00001	NM_0005	no	no	A_C	A:T>G:C	A	G		NA	NA
28824	7536380	7595615	7692297	5'UTR		0	A>G	c.1-15703	g.7595615	g.7692297	p.?	NC_00001	NC_00001	NM_0005	no	no	C_T	A:T>G:C	A	G		NA	NA

(Fig4.4) mutation database csv file format.

- Import necessary libraries: pandas, sklearn, matplotlib
- Feature selection: drop unusable columns
- Pre-processing:
    1. Remove null values
    2. Filtering data with codon 248 and 249
    3. Balance dataset (2000 cancers)
    4. Give high weight to pre-cancer class because the number of samples of pre-cancer less than number of samples of cancer
- Label encoder: to the categorical data
- Splitting data into 80%train , 20%test
- Algorithms:
    1. **ANN: Multi-layer Perceptron (MLP)** is a supervised learning algorithm that learns a function $f(.): R^M \rightarrow R^\circ$ by training on a dataset,

where M is the number of dimensions for input and O is the number of dimensions for output. Given a set of features $X = x1, x2, \ldots., xm$ and a target , it can learn a non-linear function approximator for either classification or regression. It is different from logistic regression, in that between the input and the output layer, there can be one or more non-linear layers, called hidden layers. Figure 1 shows a one hidden layer MLP with scalar output.



(Fig4.5)Neural Network

The leftmost layer, known as the input layer, consists of a set of neurons {xi|x1,x2,…,xm} representing the input features. Each neuron in the hidden layer transforms the values from the previous layer with a weighted linear summation w_(1) x_1+w_2

x_2+⋯+w_m x_m , followed by a non-linear activation function g(.):R→R.

- like the hyperbolic tan function. The output layer receives the values from the last hidden layer and transforms them into output values.

The module contains the public attributes coefs_ and intercepts_. coefs_ is a list of weight matrices, where weight matrix at index i represents the weights between layer i and layer i+1 . intercepts_ is a list of bias vectors, where the vector at index i represents the bias values added to layer i+1.

The advantages of Multi-layer Perceptron are:

• Capability to learn non-linear models.

• Capability to learn models in real-time (on-line learning) using partial_fit.

The disadvantages of Multi-layer Perceptron (MLP) include:

• MLP with hidden layers have a non-convex loss function where there exists more than one local minimum. Therefore different random weight initializations can lead to different validation accuracy.

• MLP requires tuning a number of hyperparameters such as the number of hidden neurons, layers, and iterations.

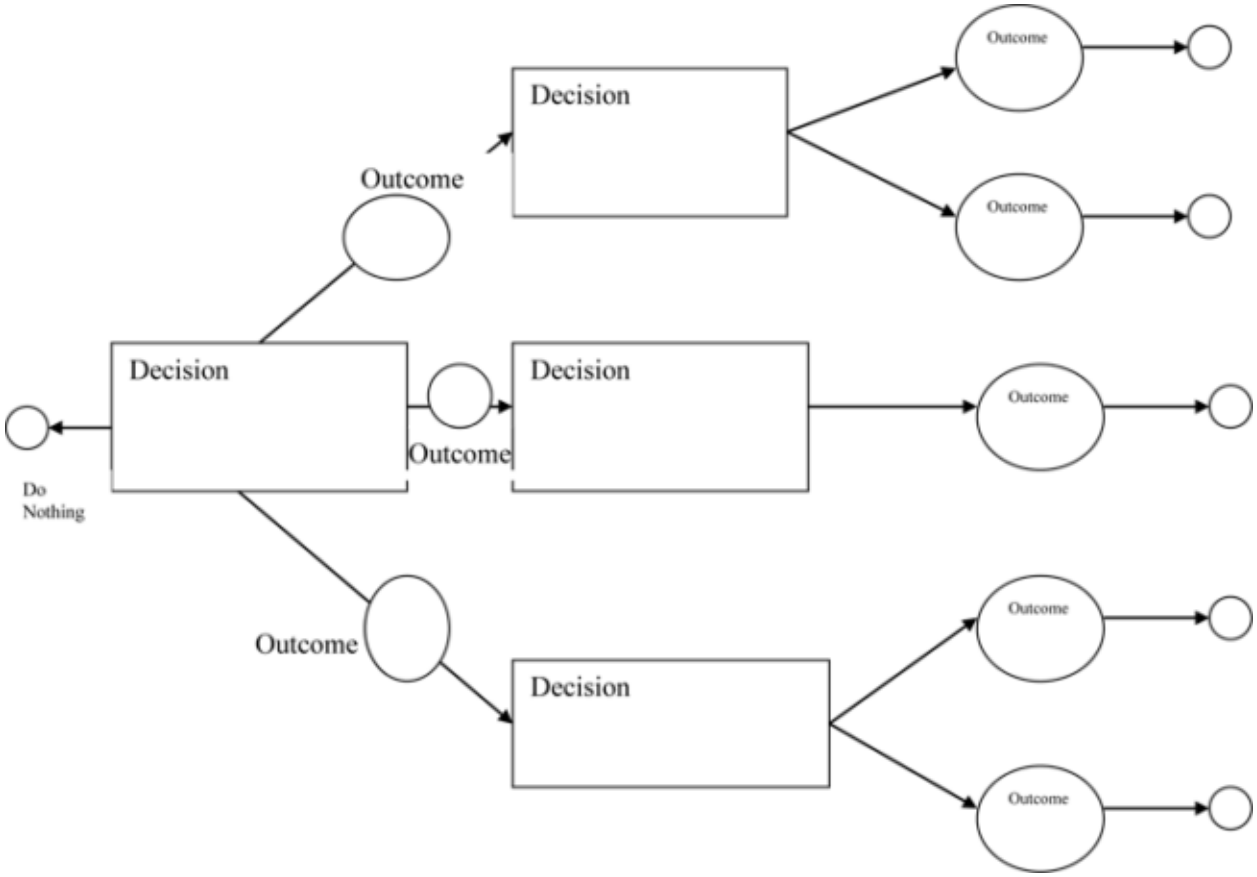
• MLP is sensitive to feature scaling.

2. Decision tree:

• Decision tree algorithm falls under the category of supervised learning. They can be used to solve both regression and classification problems.

• Decision tree uses the tree representation to solve the problem in which each leaf node corresponds to a class label and attributes are represented on the internal node of the tree.

• We can represent any boolean function on discrete attributes using the decision tree.



(Fig4.6) Decision Tree

Below are some assumptions that we made while using decision tree:

• At the beginning, we consider the whole training set as the root.

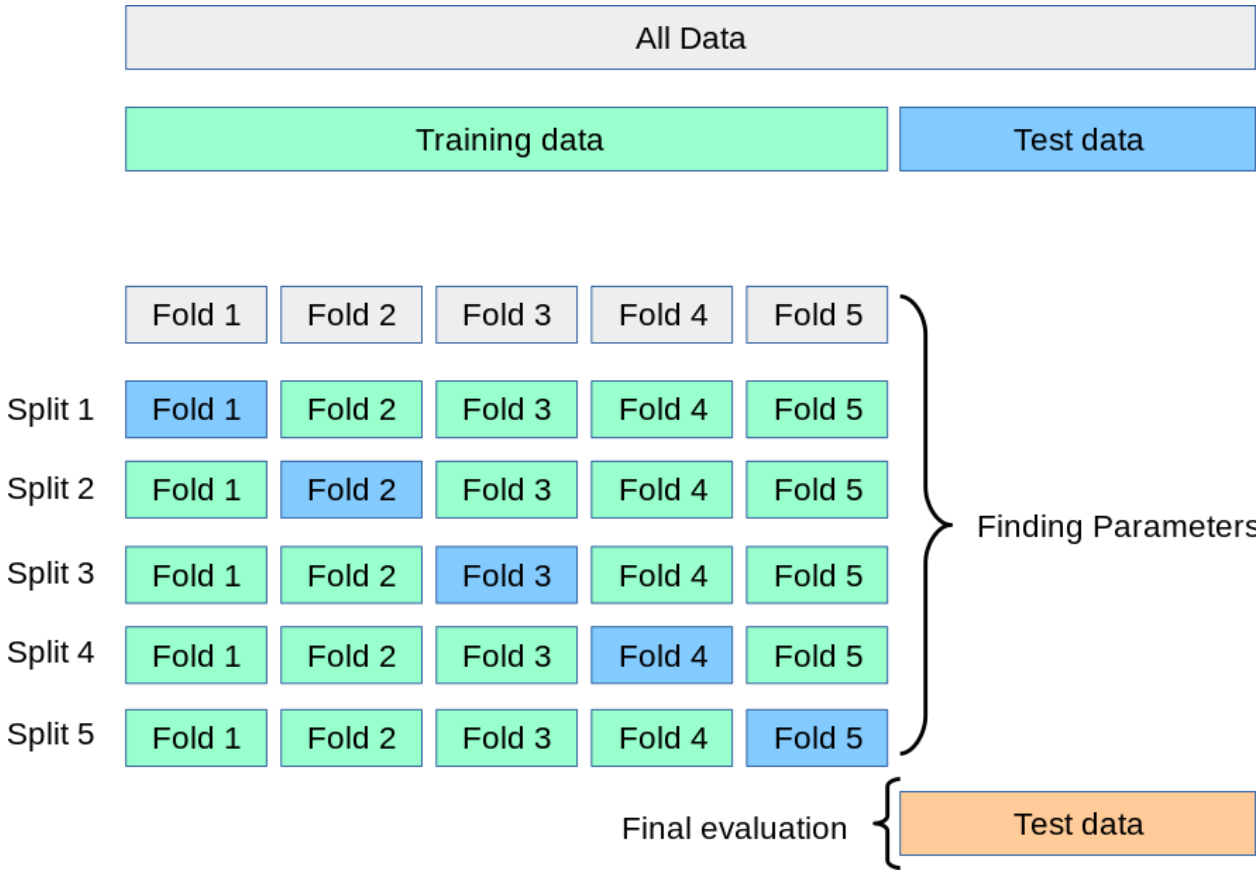
• Feature values are preferred to be categorical. If the values are continuous then they are discretized prior to building the model.

• On the basis of attribute values records are distributed recursively.

• We use statistical methods for ordering attributes as root or the internal node.

3.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.



(Fig4.7) 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.

Trails to improve the accuracy:

• Cross validation

• Balance dataset by giving the pre-cancer class high weight than cancer class.

Before balance the Training accuracy was 97.7%, Testing accuracy was 96.6%.

After balance the Training accuracy 98.5%, Testing accuracy 97.1%, so we apply this algorithm in our application.

Chapter 5

User Manual

1-open Vs code and run the program

the continue button will proceed to the main program

Home page of the system fig[5.1]

Fig[5.1]home page

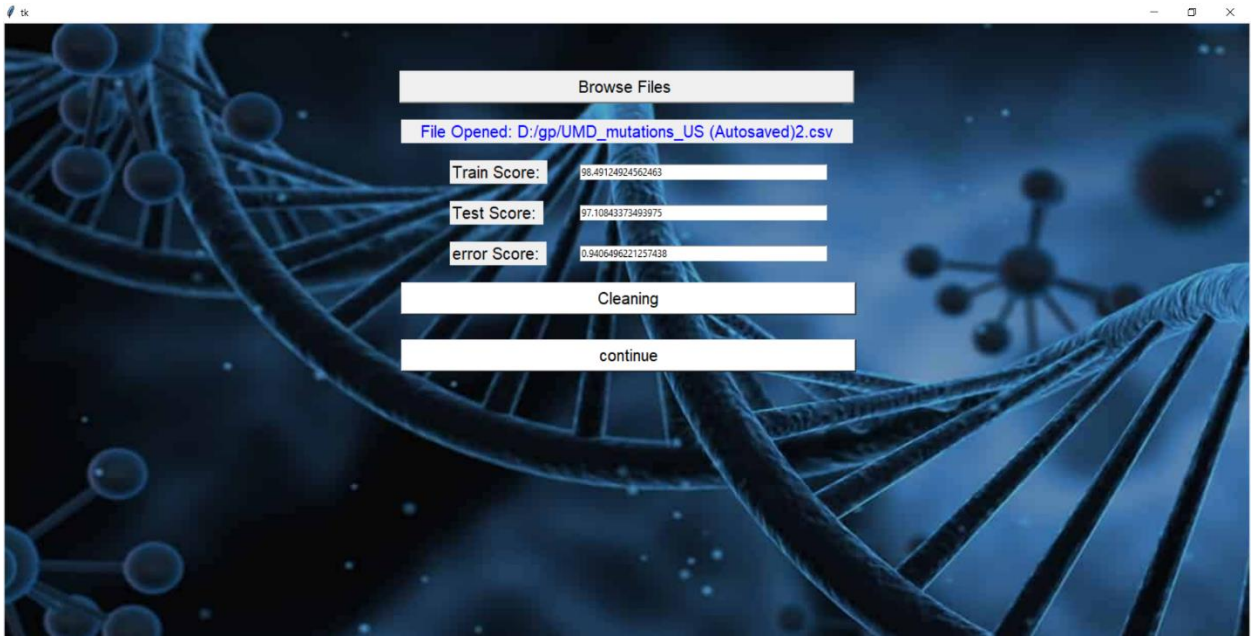
2-choose the dataset to test the accuracy
First we browse then we press cleaning button to run the model and start pre-processing.

The path of the file will appear.

The model will run and calculate train and test and error scores.

The continue button will switch the page to the testing page

Result of model training tab fig[5.2]



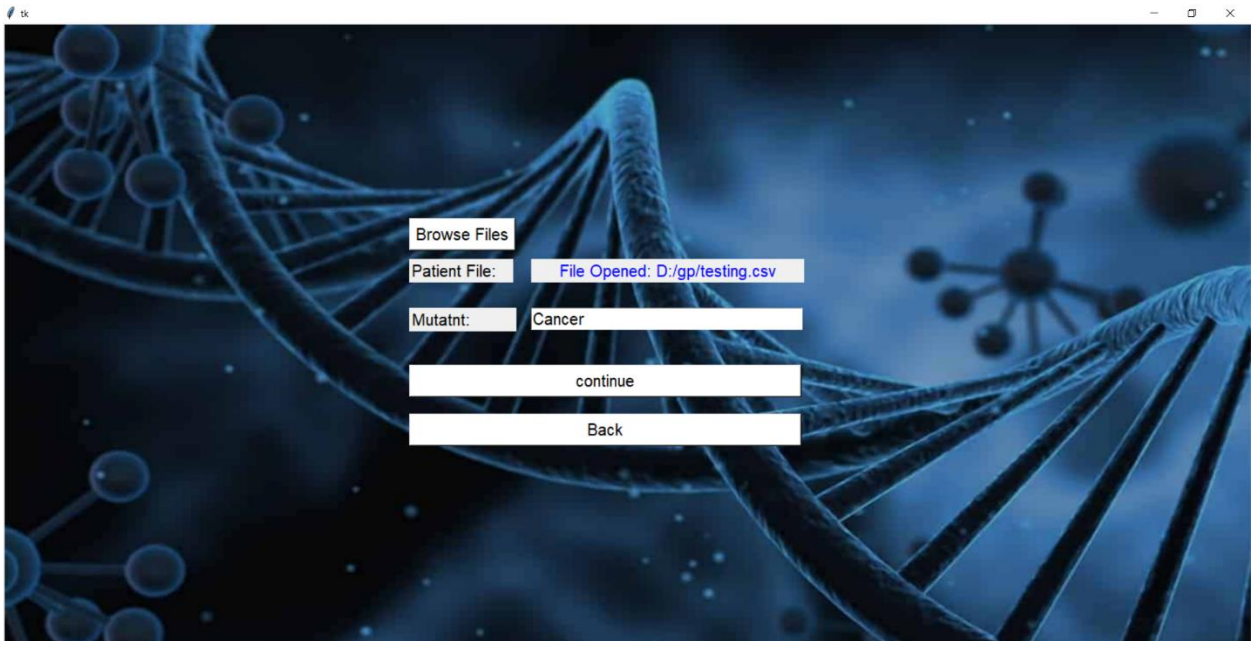
Fig[5.2] accuracy results

3-The User starts to choose the patient file and figure out if he well have cancer or pre-cancer

If we press back it will return to the testing accuracy page.

Continue button will proceed to the final page.

Testing model with un seen patient record to get cell class (cancer cell or pre-cancer cell) fig[5.3]



Fig[5.3]Testing

Last page fig[5.4]

Fig[5.4]

Chapter 6
Conclusions and Future Work

6.1 Conclusions

TP53 gene provides instructions for making a protein called tumor P53 protein, prediction pre-cancer through mutations at tumor protein P53,here we used Artificial Neural Network to predict Cancer or Pre-Cancer which is caused by 2 mutant codons which are 248 and 249 Table[6.1.1], the prediction pre-cancer must be based on collection datasets (normal TP53 gene and its tumor protein P53), as well as the TP53 database if exist which used for diagnosis and classify the type of malignant mutations and the cancer caused by the tumor protein P53 related to environment that means the prediction is more expressive for the region and will be better as this as have seen from the above impact on the results and how it could affect the environment.

Table of accuracies

ALGORITHM	CODON NUMBER	NO. FEATURES	TRAINING ACCURACY	TESTING ACCURACY	MAXIMUM DEPTH
ANN	248	101	99.8	98.9	Default
ANN	248	67	98.5	96	Default
ANN	248,249	37	97.9	94	Default
ANN	248,249	33	98	95.5	Default
ANN	248,249	30	98.8	94.2	Default
ANN	248,249	53	95.3	95.7	Default
ANN	248,249	8	95.6	94.5	Default
ANN	248,249	35	95.7	94	Default
DT	248,249	35	96.8	96.1	5
DT	248,249	35	97.4	96.1	6
DT	248,249	35	97.7	96.6	7

After enhancement:

DT→ Training accuracy 98.5%, Testing accuracy 97.1%

Here we tried ANN and Decision Tree Algorithms to predict cancer and pre-cancer caused by specific 2 mutated codons (248 & 249) of the tumor protein P53.

6.2 Future Work

An interesting aspect of this work may be explored in the future
Be to test this:

1. Identification of tissues likely to be affected by cancer by locating a precancer cell or identifying a cell that is already affected.

2. Identification of pre-cancer cell in the images datasets

3. Design an enzyme using coputer aid to repair the mutation in the damaged or affected gene.

References

[1] Kumaraswamy Naidu Chitrala, Mitzi Nagarkatti, Prakash Nagarkatti, and Suneetha Yeguvapalli1, “Analysis of the TP53 Deleterious Single Nucleotide Polymorphisms Impact on Estrogen Receptor Alpha-p53 Interaction: A Machine Learning Approach”, International journal of molecular sciences, (2019 Jun 18).

[2] Mark F Rogers, Tom R Gaunt, Colin Campbell, “Prediction of driver variants in the cancer genome via machine learning methodologies”. Oxford academic (Briefings in Bioinformatics, Volume 22, Issue 4, July 2021, bbaa250) (22 October 2020).

[3] Deeman Yousif Mahmood, Prof. Dr. Ayad Ghany Ismaeel, Prof. Dr. Abbas Hassan Taqi, “Mining Method for Cancer and Pre-Cancer Detection Caused by Mutant Codon 248 in TP53”. Periodicals of Engineering and Natural Sciences Vol. 7, No. 2, pp.522-533, (August 2019),

[4] Sayed Mohammad Ebrahim Sahraeian, Ruolin Liu, Bayo Lau, Karl Podesta, Marghoob Mohiyuddin & Hugo Y. K. Lam, “Deep convolutional neural networks for accurate somatic mutation detection”. Nature journal, (04 March 2019).

[5]Sara Saab,Hussein Ali Zalzale, Zahraa Rahal, Humam Kadara," Insights Into Lung Cancer Immune-Based Biology, Prevention, and Treatment",(Feb 2020)

[6] Wiley-Liss, Oxford, Chapter 14Mutation, Repair and Recombination ,2002

[7] Sharaf J. Malebary & Yaser Daanial Khan, "Evaluating machine learning methodologies for identification of cancer driver genes".
Nature journal Article number: 12281 (10 june 2021),

[8] Derrick E. Wood, James R. White, Andrew Georgiadis,"A machine learning approach for somatic mutation discovery". NCBI, (5 sep 2018).

[9] Nathan Wan, David Weinberg, Imran S. Haque, "Machine learning enables detection of early-stage colorectal cancer by whole-genome sequencing of plasma cell-free DNA", BMC Cancer journal, (2019).

[10] Suleyman Vural, Xiaosheng Wang & Chittibabu Guda, "Classification of breast cancer patients using somatic mutation profiles and machine learning approaches", BMC Cancer journal, (2016).

[11] Li Tai Fang, "SomaticSeq: An Ensemble and Machine Learning Method to Detect Somatic Mutations", SpringerLink, (03 March 2020).

[12] Dina Y. Mikhail, "Pre-cancer Diagnosis via TP53 Gene Mutations by Using Bioinformatics & Neural Network", 2019 International Engineering Conference (IEC) (IEEE), (23-25 June 2019).

[13] Ayad. Ghany Ismaeel, Raghad. Zuhair Yousif,” Novel Mining of Cancer via Mutation in Tumor Protein P53 using Quick Propagation Network’’

[14] Ayad Ghany Ismaeel,” New Approach for Prediction Pre-cancer via Detecting Mutated in Tumor Protein P53”, (October 2013)

[15] International Journal of Computer Science and Electronics Engineering (IJCSEE) Volume 3, Issue 2 (2015) ISSN 2320–4028

[16]Minetta C Liu,Edward P Gelmann,” P53 gene mutations: case study of a clinical marker for solid tumors”,(Jun 2002)

[17]Laurent-Puig P, Béroud C, Soussi T. APC gene: database of germline and somatic mutations in human tumors and cell lines. Nucleic Acids Res 1998

[18] Béroud C, Soussi T. p53 and APC gene mutations: software and databases.Nucleic Acids Res 1997

[19] Soussi T, Béroud C. Assessing TP53 status in human tumours to evaluate clinical outcome. Nat Rev Cancer 2001

[20] Hamroun D, Kato S, Ishioka C, Claustres M, Béroud C, Soussi T. The UMD TP53 database and website: update and revisions. Hum Mutat. 2006