# AI-Driven Strategies for Analyzing TP53 Mutations: Detection and Structural Implications

Aruna Arumugam, IAENG, Member, Mohamed Divan Masood

*Abstract* - **Identifying the DNA sequence mutation that causes cancer is challenging in precision oncology. Subsequently, many advanced algorithms and techniques are available now to trace the cancer cell, but systematic approaches to optimize the huge dataset of DNA sequence structure are few. The mutant p53 protein, which is present in chromosome 17, will turn against apoptosis (programmed cell death) or cell cycle arrest (p21) and has a high possibility of oncogenesis (malignant cancer). In this research work, we customized and employed the Rabin-Karp algorithm to search the DNA mutation sequence string and then feed the output obtained to train the model using deep learning techniques. This helps us to analyze and observe the change in DNA structure patterns that cause mutation, especially in the p53 protein, which is the major responsible for most of the cancer variants.**

*Index Terms* - **DNA sequence mutation; Rabin-Karp; gene structure; deep learning computational model; tumor suppress gene; oncology**

## I. Introduction

ONE of the life-threatening diseases is cancer. Abnormal or uncontrollable growth of the human body's cells is the reason for cancer. Cancer can be of two types i) benign – slow in growth and doesn't spread ii) malignant – fast in growth and spread to the other organs of the body. There are so many reasons for cancer cells to grow some of them are damaged DNA sequences attained from environmental factors, food habits, tobacco, alcohol, lifestyle, error during the life cycle process of cell division, or may be inherited from the parents or grandparents. Cells can communicate with each other through signals[1]. In general, cells are grown only when they receive their command or indication to grow but these cancer cells will grow in the absence of the signal. Some cancer cells replicate double the count of normal cells. By default, our human body is sanctified by the immune system. Sometimes these tumor cells will convince or put a mask on from the immune cells to keep the cancer cell alive instead of killing them. Initially, cancer can form in any organ of the body and that is called primary cancer[2]. The blood vessel will carry the affected or damaged cell to other parts of the body and spread cancer. This dangerous sort of tumor is called metastasis.

The change in the DNA sequence structure will contribute to cancer. These tend to affect three main types of genes which will fight against cancer cells. 1) Tumor suppressor gene - has control over cell growth and division. Some alterations in the tumor suppressor gene lead to developing the damaged or cancer cells uncontrollably. 2) Proto-oncogene - allows growing the mutated cells that are not intended or supposed to grow. 3) DNA repair gene - the job of this gene is to rectify the bug in DNA damage. If certain alteration happens with this gene will add addition mutation in chromosomes like deletion and duplication in quick manner. The detailed explanations about the cell and mutations are carried out in the following sessions.

Manuscript received August 1, 2024; revised March 3, 2025.

Aruna Arumugam is a PhD candidate in the Department of Computer Application at B. S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, Tamilnadu, India. (phone: +91 7358743588, e-mail: arunanet23@gmail.com)

Mohamed Divan Masood is an Assistant Professor in the Department of Computer Application at B. S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, Tamilnadu, India. (corresponding author to provide phone: +91 9790404629, e-mail: divan@crescent.education)

## II. The Stages of the Cell Cycle

A cell is the beginning of everything. Every human body is made up of not less than 60 trillion cells. It's the smallest unit that makes up all living or breathing organisms that can live on its own. The nucleus, the cell membrane, and the cytoplasm are the three major parts of the cell structure as shown in Figure 1.

Most of the cell's DNA is resides inside the nucleus structure and this is the place where most of the RNA also made. The cytoplasm is made up of fluid that consists of tiny cell parts including the Endoplasmic reticulum, Mitochondria, and the Golgi complex. The cytoplasm consists of jelly-like fluid. The endoplasmic reticulum (ER) is the one who is responsible for transporting the molecules either inside or outside of the cell to their particular destinations. The molecules cultivated by the endoplasmic reticulum (ER) are transported out of the cell by the Golgi complex. The mitochondria help to convert the energy from food to the form the cell can use. The nucleus will give the command to the cell to grow, divide or die. It is encircled by a membrane called the nuclear envelope. This nucleus envelope helps to discrete the nucleus from the residue of the cell and safeguard the DNA. A cell cycle is the sequence of events that is responsible for cell growth and cell division [3]. Interphase is the time when cell growth takes place, replicates or duplication of its gene structure, and make the cell ready for duplicate or division. There are four phases in life cycle of cell.

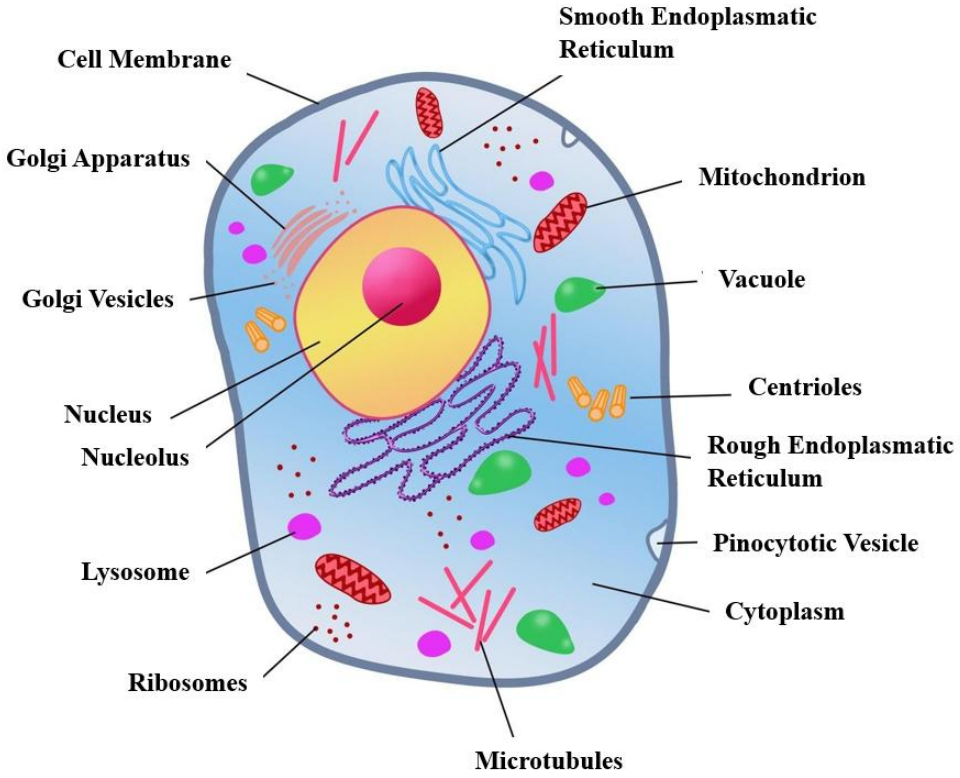



Fig 1. Parts of a human cell. Nucleus with centrally positioned chromosomes (DNA) containing genetic material.

1) G1 Phase (gap phase): Initiate the cell division and moves to the next phase S.
2) S Phase:(synthesis phase): DNA replication of cells will happen here. Each gene will have sister chromatids now.
3) G2 Phase (gap phase): It assemble the cytoplasmic needed for mitosis and cytokinesis.
4) M Phase (mitosis/meiosis phase): Cell division with two (Mitosis) or four (Meiosis) daughter's cells will take place here.

## III. Fundamental of the Chromosome Biology

Each cell nucleus consists of spindle fibers and thread-like structures called chromosomes. The two important portions of chromosomes are proteins and DNA. Histone protein is the one that helps to maintain the formation of the DNA (Deoxyribonucleic Acid). Each chromosome has three parts. The center joining part is called the centromere. The edge of each chromosome is called the telomere. The TTAGGG (thymine (T), adenine (A), guanine (G)) is the telomere sequence which is repeated over 14000 base pairs. This plays an important role in keeping the genomic sequence by safeguarding it from fusion, degradation and recombination. An individual piece or structure of a chromosome is called a chromatid. The upper part, which is small in shape is called the P arm. The lower part, which is long in shape is called the q arm. The distance from the centromere to the P and Q arm will determine the shape and the characteristics of the specific genetic [4].

All living organisms have various counts and shapes in chromosomes. A cat is comprised of 19 couples of chromosomes with the individual count as 38. A dog is comprised of 39 couples of chromosomes with the individual count as 78. Our human body is built up of 23 couples of chromosomes which is equal to 46 individuals. The first 22 couples are named autosome. This couple is the formation of one copy chromosome from the mother (female parent) and one copy chromosome from the father (male parent) as an inheritance. The last pair is the twenty third chromosome named as the sex chromosome and is used to determine the gender of the child. If the combination of X is from the mother and X is from the father then the child is a female (XX). If the pairs are created as X from the mother and Y from the father then the child is male (XY). The possibility of chromosome disorder is the changes in the count of a chromosome or the structure. Half (23) of our gene structures are from the mother, and the other half (23) are from the father. It raises the risk of inherited or heredity genetic disorders [5]. The disorder can be by birth itself, or it might get developed over time.

1) Numeric abnormalities

The mislaying or inclusion of the entire chromosome, which can attack hundreds to thousands of gene structures, is deadly. Some are the fatal diseases caused by numeric abnormalities of chromosomes are Patau's, Edward's, Down, and Klinefelter's syndrome as shown in the following Figure 2 [7]. Our human bodies have the in-built mechanism to rectify automatically the small or few alterations in gene structure because of the immune systems.

2) Structural abnormalities

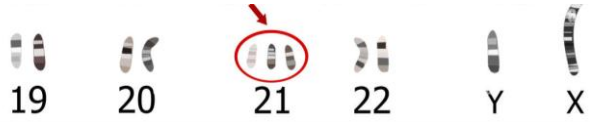
Where a large count of DNA structures is deleted or multiple copies take place. The foremost causes of cancer are structural anomalies, consisting of oncogene translocations (e.g., BCR-ABL in continual myeloid leukemia). Important genes can be deleted or duplicated, main to developmental or inherited problems (e.g., trisomy 21 reasons Down syndrome). Abnormal DNA structure can motive gene expression and protein feature to be disrupted, that could

bring about mobile dysfunction or demise. Comprehending structural anomalies gives valuable statistics for targeted treatments, like medicines that concentrate on translocation-brought on fusion proteins. These can take several forms as shown in Figure 3.

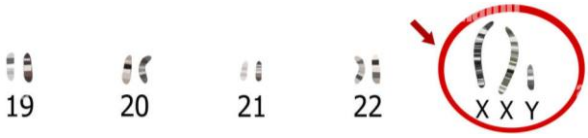
1) Deletion - some of the chromosomes will be lost.
2) Duplication - Multiple copies occurred in duplication.
3) Inversion - DNA pattern copies in a reverse way.
4) Ring chromosome - Chromosomes change the shape vertically to ring.
5) Reciprocal - Patterns from the two various chromosomes are exchanged.
6) Robertsonian - a complete chromosome will get couple with other.


(a) Patau's syndrome.


(b) Edward's syndrome.




(c) Down syndrome.




(d) Klinefelter's syndrome.

Fig 2. Possibilities of numeric abnormalities.

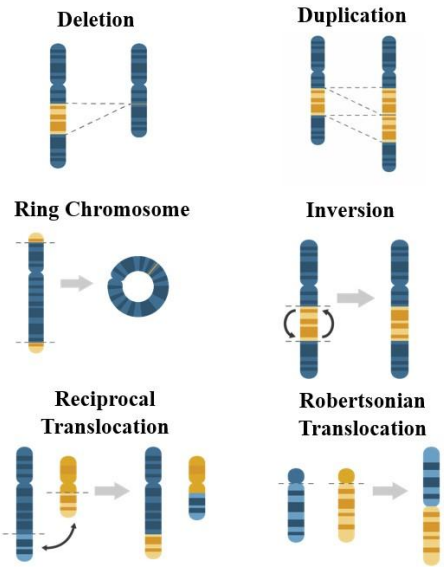



Fig 3. Possibilities of structural abnormalities.

## IV. The Sequence of the DNA Structure

The inherited material of the human body is the DNA or Deoxyribonucleic Acid. The DNA is the blueprint for all breathing organisms. The majority of the DNA is present in the nucleus of the cell and, a few of the DNA is also present in the mitochondrial which can be called mitochondrial DNA or mtDNA. This mtDNA helps to convert food into energy which is required for the cell. It is made up of four nucleotides called Adenine (A), Cytosine (C), Guanine (G), and Thymine (T)[8]. These four chemical bases are called building bricks of DNA.

Our human body comprises more than three billion base pairs of DNAs. These cells can hold the information or signals to grow or stop the growing process of the organs at some point. For example, the normal kidney is 10 to 12 cm in size. The kidney starts the growth development from birth itself. But, once it reaches its specific maturation it should stop growing further, or else that also leads to the health issues. All these development processes are maintained, monitored, and controlled by the cells of our body.

In the base pair of the DNA, Adenine is consistently paired with Thymine and Cytosine with Guanine (A with T, G with C). The sequence or the order of the base pair may vary. The number of times adenine will have the same amount of thymine. Similarly, the count of cytosine matches the count of guanine. The boundary of the DNA structure is comprised of sugar and phosphate.

The two folded helix will have the appearance of a twisted ladder [9]. One with the end of 5'(five prime ends) and the other with 3' (three prime ends). This formation looks anti-parallel. That is, one DNA strand with a 5' end is collateral to the 3' end with the other strand.

### *A. Transcription and translation in molecular biology*

An immense sequence of the human genome that is non-coded and transcribed is called introns. Our human body consists of 98% of this non-coded DNA. These non-coded are detached from the nucleus before the messenger RNA (mRNA) moves to the cytoplasm. The DNA sequence is called exons when they are coded. Our human body consists of 2% of this coded DNA and, it is responsible for encoding proteins.

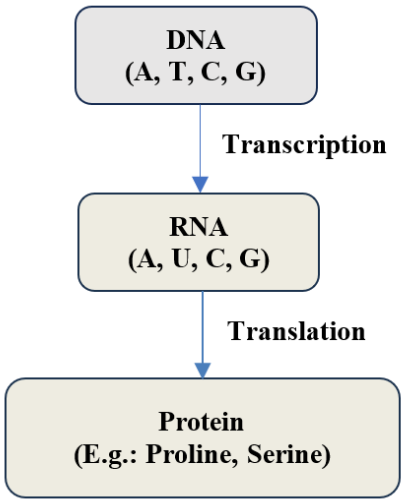



Fig 4. Conversion process of DNA to RNA and to protein.

1) Transcription - It is a procedure by which the DNA is transcribed or copied to messenger RNA (mRNA) that holds the details for the synthesis of the protein.
2) Translation - After the transcription, the RNA is used to construct proteins is called translation. This cannot be done directly by mRNA but with the help of transfer RNA (tRNA) as shown in Figure 4.

### *B. Example of the Genetic Code*

(A, T, C, G) – 4 bases of DNA (Adenine, Thymine, Cytosine, Guanine).
(A, C, G, U) – 4 bases of RNA (Adenine, Cytosine, Guanine, Instead of Thymine in RNA we have Uracil).
One strand of DNA sequence. Let's say "A" 5'TCGTCGACGATGATCATCGGCTACTCGA3'.
This strand will form the antiparallel DNA strand "B" as follows,
3'AGCAGCTGCTACTAGTAGCCGATGAGCT5'.
The messenger RNA (mRNA) transcription of strand "A" is as follows, (refer amino acid table).
5'UCGAGUAGCCGAUGAUCAUCGUCGACGA3'.
Because the four bases of RNA are A, G, C, and, U. The sequence of amino acid for the above mRNA is as follows,
Ser-Ser-Ser-Arg-STOP (Translation).

## V. DNA MUTATION AND THE SOURCES OF THE CANCER

Our human gene holds all information that is required to build our body. For example, blue eyes, brown skin, and thick hair. It decides the cells of what kind, and when they should grow, divide, and die. During cell division, the gene grabs all the mistakes that happen in DNA sequence copying. That mistake is called a mutation. One can have a mutation of a gene all through life. But by default, our gene will act like a guardian and pick up all the errors caused by the mutation and correctly arrange the DNA sequence. But sometimes the mutated cells hide behind the normal cells and started to replicate further in a faster way. These leads to huge issues like transporting the cancer cell from one part of the organ to other parts of the body through blood vessels.
The followings are the reasons the gene gets mutated at the time of cell division.

1. Oncogene - It is one of the genes that instruct the cell to multiply and divide. If the faulty cells are divided and make a huge number of copies in an uncontrolled way, then it leads to cancer.
2. Tumor suppressor gene - It is called the guardian of the genome [10]. It has the control to repair and rescue the mutant cell. If it cannot able to repair the cells it will destroy or kill them and prevent them from further spreading or dividing. The p53 protein present in chromosome 17 is one of the tumor suppressor genes. Apart from this, there is p21, PTEN, p16, and, retinoblastoma protein (pRb). Most cancer causes because by the mutant p53. Once the tumor suppressor gene gets mutated its losses its control over cells to instruct or be idle without protecting the cells from damages.
3. DNA repair gene - According to the gene, our cells have the protein whose task is to repair the DNA-damaged cells. The cell can repair itself but that is not sufficient for increased damage. If the gene stops producing the proteins that repair the cell, then it forms the cancer cells.
4. Self-destroyed gene - When a cell is too damaged or, became old, our gene has the programmed cell death function which is called apoptosis. If a gene detects something wrong during the process of cell division it instructs the cell to kill itself.

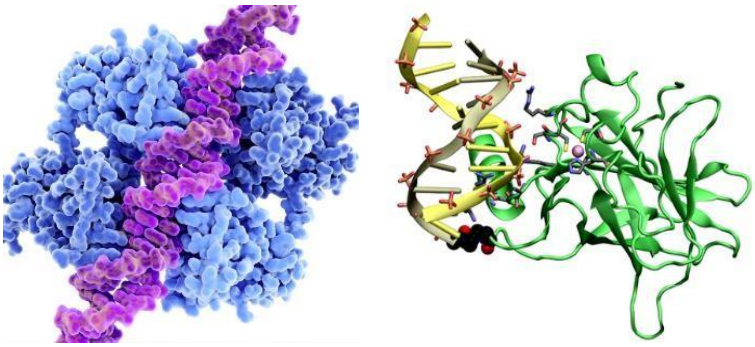
### *A. The guardian of the genome*

The proto-oncogene is the beneficial cell cycle regulator which helps to prevent oncogene. An oncogene is the cell growth and cell division of the cancerous cell. If the proto-oncogene is mutated in a certain way, it leads to the development or growth of the cancer cell in an uncontrolled manner. Proto-oncogene helps to code the proteins and enable the mitogenic signals. To control the mistakes that happen in cell growth our gene has endowed with the tumor suppressor gene called p53, p16, p21, PTEN, and pRb. These tumor suppressor genes [11] help to control the abnormal and uncontrollable growth of the cells. That is the reason these proteins are called guardians of the genome [12].

1. Gene p16 - p16 is the multiple tumor suppressor and is known as cyclin dependent kinase inhibitor 2A (CDKN2A). Located in chromosome 9. The protein helps slow down the cell division process during the cell cycle stage from the G1 phase to the S phase and prevents cancer cells. If the protein is mutated it loses its function and becomes the cause of many variants of cancers.
2. Gene p21 - it is also known as cyclin-dependent kinase inhibitor 1 (CDKN1A) present in chromosome 6. This protein is involved in the activity of p53 associated with cell cycle arrest in the process of protecting against DNA damage.
3. Gene PTEN - It instructs for creating an enzyme that is present in the tissues of the human body. The protein is known as Phosphatase and tensin homolog (PTEN). This protein works the same way as p16 and slows down or stops the process of cell division rapidly in a faulty manner. Located in chromosome 10.
4. Gene pRb - It is located in chromosome 13. Mutation in this protein leads to the majority of cancers. The protein is known as retinoblastoma protein (pRb) and it is the proto- oncogene. Helps to prevent extravagant cell growth.
5. Gene p53 - More than 60% of the cancer are caused by the mutant p53 protein. This tumor suppressor protein prevents the formation of cancerous cells [13]. That is the reason it is named "the guardian of the genome". The location of the protein is in chromosome 17. It helps to encode the proteins that are in the DNA domain binding and manage the expression of the gene to prevent genome mutation. The p53 gene encodes up to 15 protein variations or isoforms in full length proteins. Protein variation or isoform is from the same gene family but will yield differences in genetics by alternative splicing. Figure 5(a) shows p53 tumor suppressor protein, and (b) shows in what way the single piece of p53 protein is communicating with the fragment of DNA strands.

### *B. Structure of the P53 gene*

The p53 protein is situated in chromosome 17 of the human body. It has 11 exons and, 10 introns within the structure [14]. It is comprised of 393 amino acids and, consists of five domains. 1. From (1-60) - N terminal transactivation domain,



(a) TP53 protein. (b) Interaction of a DNA fragment with a single p53 unit.

Fig 5. Interaction of tumor suppressor protein (TP53).

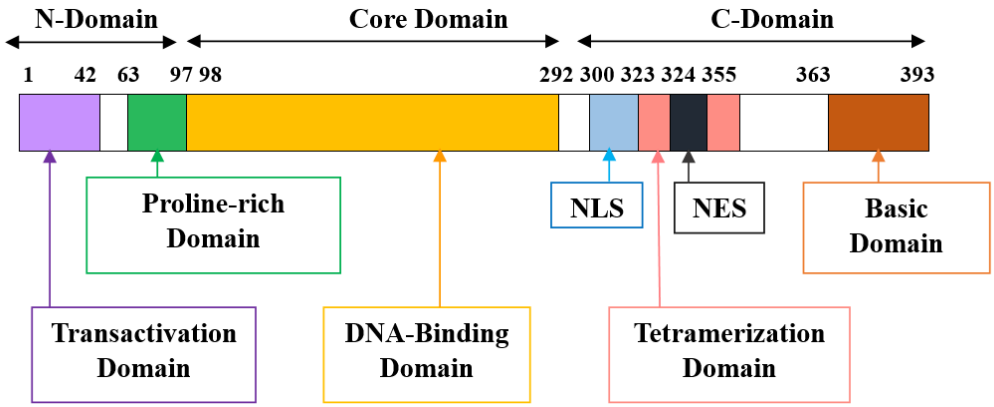



Fig 6. Structure of the p53 gene on chromosome 17.

2. From (60-100) - proline-rich domain, 3. From (100-300) - DNA binding domain, 4. From (323-355) – tetramerization domain, and the 5. From (363-393) - C terminal basic domain as shown in the Figure 6.

Nuclear Localization Signal (NLS) and the Nuclear Export Signals (NES) are the signals that helped to carry out the nucleocytoplasmic transport by the p53 protein [15]. The protein p53 plays a vital role in cell growth and cell programmed death [16]. If the p53 got defeated by some facts then there is a possibility of abnormal and uncontrollable growth of cells that, consequence outcome in cancer. More than 60% of the cancer formations are because of the mutant p53. In the typical cell, the level of the p53 protein is low. It can be triggered only by the sensation of 8 signals from DNA stress and repair.

It involves three majority roles. The first one is the cell growth arrest when it is needed. Some of the examples are p21, 14-3-3s, and Gadd45. The next one is repairing the DNA. An example of this is p53R2. The third one is programmed cell death [17]. Some examples of this are PUMA, Noxa, and BAX. The MDM2 (mouse double minute 2) is the negative regulatory factor and phosphoprotein that joins the p53. If the level of MDM2 increases, it suppresses the functions of the protein p53 and increases the risk of cancer.

## VI. Computational Model

*Methods*

### A. *Rabin-Karp Algorithm for DNA sequence representation*

The Rabin-Karp algorithm uses a hash function to search for and match patterns in text [18]. In the first stage, it doesn't go over every character; instead, it filters the ones that don't match before doing the comparison. A string of characters is selected and its potential to contain the necessary text is examined. Character matching is done if the possibility is discovered. To examine the method using the subsequent steps:

*Step 1:* Assign the value for nucleotide bases A=1, C=2, G=3, T=4.

*Step 2:* The non-mutant DNA string be

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| CGT | GCT | TTC | CAC | GAC | GGT | GAC | ACG |

The Pattern supposed to be at location 5 is *GAC*.

*Step 3:* To find the hash value for pattern(p) is

$$\sum(v * dm - 1) \quad (1)$$

Where, d =30 is the total length of the string, m =3 is the pattern size to search.

$$= ((3 * 30^2) + (1 * 30^1) + (2 * 30^0))$$

*H(GAC) at location 5* = 2732

*Step 4:* In case the given DNA string with pattern at location 5 be like this as follows

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| CGT | GCT | TTC | CAC | GCC | GGT | GAC | ACG |

Now the hash value = ((3*302) + (2*301) + (2*300))

*H(GCC) at location 5*= 2762

*Step 5:* Compare both the hash value at same location 5. The value and pattern string differs from the original DNA sequence string.

By using this Rabin-Karp algorithm[19], we can efficiently detect the mutant, insertion, deletion, frameshift pattern in P53 DNA sequence and train the model.

### *B. One-Dimensional Conventional Neural Network (1D-CNN)*

1D convolution layers are very good at identifying specific patterns within a series. They use the same transformation on each patch, so a pattern that is learned at one point in a series may be recognized at a different place later on. 1D convolutional networks [20] with DNA motif identification are made possible by this feature.

For example, the network should be able to learn motifs or DNA fragments up to three characters in length while processing DNA sequences using 1D convolutions with a window size of three. Furthermore, it is able to identify these words in any input sequence context.

The one-dimensional CNN was made up of the subsequent layers:

*Step 1:* Input layer: Our model uses a 1D vector as its input, which is a vector with a size of 1 × 100 that was produced using the Rabin-Karp Pattern Matching algorithm.

*Step 2:* Convolutional layer: Features encoded in the 1D input vector are derived by building a convolution kernel using a 1D convolutional layer [21]. Using a sliding window (window size 3), the convolutional layer carefully scans the input, turning the values into representative values. By extracting useful features from small portions of the input data, this procedure aids in the conservation of the dimensional relationship between the numerical values in the vectors. We used a kernel size of 3 to extract more information because our input size was small. As shown in Figure 7 we employ 1D convolutions to extract local 1D patches or subsequence from sequences.

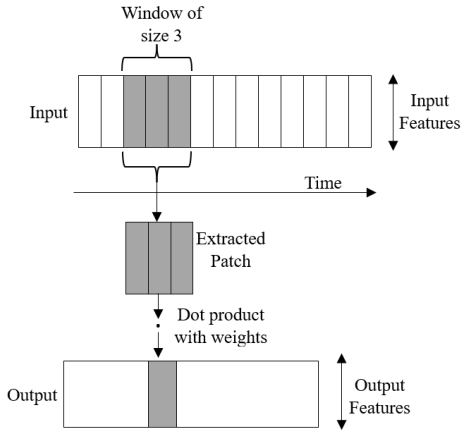



Fig 7. 1D Convolutions for DNA sequence extraction.

*Step 3:* Rectified Linear Unit (ReLU): Following each convolution operation, a new non-linear operation is presented. Its goal is to carry out non-linear functions in our CNN and improve our model's comprehension of the data. The following is the output function of ReLU:

$$f(x) = \max(0, x) \tag{2}$$

where x is a neural network's input count.

*Step 4:* The pooling layer is typically incorporated into the convolutional layers to minimize computation for subsequent layers [22].

*Step 5:* Dropout layer: A method designed to help improve the model's performance while also preventing overfitting.

*Step 6:* Flatten layer: A layer aids in the vectorization of the input matrix.

*Step 7:* Fully connected layer: typically added by the deep network's final stage. If every node in the network is connected to every other node in the network, then the layer is fully connected. It was a binary classification because our challenge is to distinguish between mutant and non-mutant sequences in the p53 DNA sequence. Consequently, our output has two nodes in total at the end.

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{k=1}^{K} e^{z_k}} \tag{3}$$

where $i^{th}$ class is the forecast probability from sample vector x, $\sigma(z)_i$ is real values in the range (0, 1), and z is the input vector with K-dimensional vector. Softmax had to be inserted in order to calculate the probability of each potential outcome [23].

Pseudocode:

```
>> Input Handling
>> Input_string ← get input from the user
```

Compare Input String with Stored DNA String:

```
>> for i from 0 to length of dna_string - 3:
>> substring ← dna_string [i to i+2]
>> if substring = input_string [i to i+2]:
>> mutation_location ← i + 1
>> mutation_value ← convert (input_string [i to i+2])
>> print "Mutation found at position " +
    mutation_location + ". Mutation value: " +
    mutation_value
```

Rabin-Karp Algorithm:

```
>> function rabin_karp (dna_string, pattern,
    mutation_value):
>> dna_mapping ← {'A': 1, 'C': 2, 'G': 3, 'T': 4}
>> numerical_dna ← convert dna_string using
    dna_mapping
>> numerical_pattern ← convert pattern using
   dna_mapping
>> dna_hash ← sum of first length(pattern) elements of
    numerical_dna
>> pattern_hash ← sum of numerical_pattern
   matches ← []
>> for i from 0 to length of dna_string - length of pattern
    + 1:
>> if dna_hash = pattern_hash and numerical_dna [i to
    i+length(pattern)-1] = numerical_pattern: append i to
    matches
>> if i < length of dna_string - length of pattern:
    dna_hash ← dna_hash - numerical_dna[i] +
    numerical_dna [i + length(pattern)]
>> return matches
```

Identify Mutations:

```
>> mutations ← rabin_karp (dna_string, pattern,
    mutation_value)
```

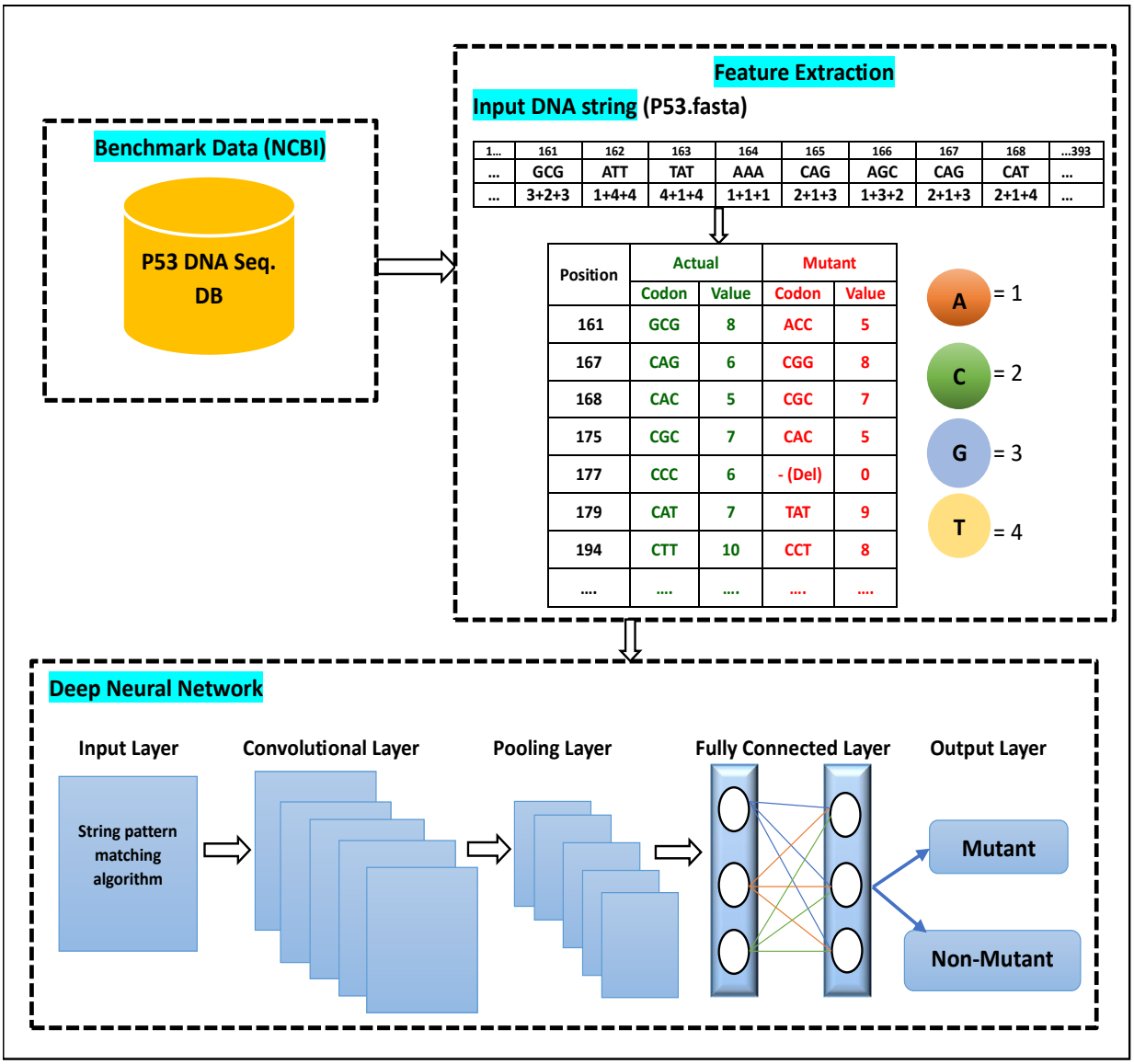



Fig 8. Architecture diagram of the proposed model (Rabin-Karp with Convolutional Neural Network-CNN).

An entire flowchart of our methodology is shown in Figure 8 under the control of a specifically designed pipeline. The following subsections will cover each of this indicated pipeline's experimental phases in detail.

The below figure shows that the location and weight of the normal and mutant codons were identified using a hash table after the model had been developed by applying the Rabin-Karp algorithm[18] on the benchmark dataset (the p53 DNA sequence from the NCBI). Next, in order to learn these feature extraction techniques and categorize the DNA sequence as either mutant or non-mutant, a deep network was constructed[24].

For implementation, the dataset should be pre-processed first. The input DNA sequence file should be a FASTA file. It requires the following four steps to perform the analysis.

1. The reference genome (DNA sequence of the p53 protein) can be downloaded in FASTA file format (https://www.ncbi.nlm.nih.gov) [23] from the National Centre for Biotechnology Information (NCBI).
2. Data Preprocessing - Remove any characters (A, T, C, G) that are not included in the DNA alphabet. To prevent inconsistencies during analysis, convert sequences to a standard format (upper/lower case). To improve the confidence of downstream analysis, remove sequences that fail to meet quality thresholds or have low-quality scores. Identify and extract pertinent sequence features, such as domains, motifs, or other patterns with biological importance [25]. To encode the DNA sequence, convert sequences to representations of numbers (K-Mer). 3)
3. Analyzed and identify the list of possible mutations in the p53 protein that leads to cancer by applying the Rabin-Karp algorithm in the DNA sequence and create the hash values to train the model.
4. For execution python and all its essential libraries (bio.seq, biopython, SeqIO, scikit-bio, pandas-bio, DNAdigest, pydna) software packages should be installed. By loading the input DNA sequence file, the model can able to accurately predict the sequence is mutant or non-mutant [26].

Following Table I presents the key hyperparameters chosen for optimizing the model's performance. These parameters were fine-tuned to enhance generalization and improve TP53 binding site prediction accuracy.

TABLE I. SELECTED HYPERPARAMETERS FOR THE MODEL'S TUNING

| Parameters | Range | Stepsize | Optimal |
|---|---|---|---|
| Learning Rate (lr) | 0.05–0.25 | 0.05 | 0.1 |
| Dimension (Dim) | 50–500 | 25 | 100 |
| Window Size (ws) | 1–10 | 1 | 5 |
| Epoch (Nos. of Iteration) | 25–500 | 25 | 100 |
| Loss (Function) | [ns, hs, softmax] | - | softmax |
| Early Stopping Patience | 3 to 10 epochs | ±1 epoch | 5 epochs |
| Number of Conv Layers | 1-10 layers | ±1 layer | 3-5 layers |
| Activation Function | ReLU, tanh, sigmoid | N/A | ReLU |
| Optimizer | Adam, SGD, RMSprop, Adagrad | N/A | Adam |
| Regularization (L2/L1) | 0.0 to 0.1 | ±0.001 | L2 = 0.001 |
| Fully Connected Layers | 64, 128, 256 neurons | Doubling | 128 neurons |
| Batch Size | 16, 32, 64, 128 | Doubling | 32-64 |
| Dropout Rate | 0.0 to 0.5 | ±0.1 | 0.3 |

The following criteria are included for measuring the performance of the models: Sensitivity (Sens), Specificity (Spec), Accuracy (Acc), and Matthews Correlation Coefficient (MCC). False Positive (FP) and False Negative (FN) indicate the associated amounts of incorrectly classified mutant and non-mutant, whilst True Positive (TP) and True Negative (TN) represent the numbers of correctly predicted mutant and non-mutant, respectively [28].

$$Accuracy = 1 - \frac{FN+FP}{(TP+FN)+(TN+FP)}, 0 \leq Acc \leq 1 \quad (4)$$

$$Specificity = 1 - \frac{FP}{TN+FP}, 0 \leq Spec \leq 1 \quad (5)$$

$$Sensitvity = 1 - \frac{FN}{(TP+FN)}, 0 \leq Sens \leq 1 \quad (6)$$

$$MCC = \frac{1-(\frac{FN}{TP+FN}+\frac{FP}{(TN+FP)})}{\sqrt{((1+\frac{FP-FN}{TP+FN})(1+\frac{FN-FP}{TN+FP})}} \quad (7)$$

## VII. COMPARISON THE PERFORMANCE RESULTS BETWEEN PROPOSED METHOD AND THE EXISTING METHODS

As previously mentioned, our model combines deep convolutional neural networks with the Rabin Karp algorithm to create string-matching patterns. It performs admirably. In order to be fair, we must contrast our suggested approach with other earlier studies on the classification of mutants. Table II displays the results, and to emphasize the importance of each metric, we have highlighted the highest values. Then, we noticed that in both layer classifications, our method performs better than other predictors in every metric (sensitivity, specificity, accuracy, and MCC). An additional enhancement is that our methodology can be utilized for whole genome sequences instead of just brief segments. To create a vector with a fixed length, all sequences of varying lengths will be trained [29]. Any type of input is helpful.

Even though there were a few computational tools available at this point [30], the performance outcomes still need to be improved. This research work introduces a novel hybrid system to identify mutant p53 DNA sequences using a combination of string-matching algorithms and deep learning. Unlike other approaches, this one allows us to generate the hidden information of DNA sequences. With an accuracy of 85.41%, it was discovered that the suggested method could distinguish between mutant and non-mutant DNA sequences and their strengths. The remaining measurement metrics, including MCC, sensitivity, and specificity, all achieved improved outcomes. In comparison to other cutting-edge predictors [31] for the same problem as well as the data set, our suggested approach has shown enhancement in all metrics by roughly 1-4 percent. Additionally, it can serve as a foundation for future research aimed at deciphering the linguistic context of DNA sequences [32]. The following Table II presents the comparative results of models.

TABLE II. COMPARISON WITH PREVIOUS PREDICTORS ON THE SAME DATASET

| Predictor | Acc | Spec | Sens | MCC |
|---|---|---|---|---|
| Rabin-Karp – CNN (Our Model) | **85.41** | **88.67** | **83.76** | **0.701** |
| K-mer - CNN[24] | 81.37 | 84.89 | 83.13 | 0.663 |
| CNN -Bi LSTM[25] | 82.02 | **89.02** | 79.1 | 0.632 |
| One Hot encoding - Random Forest[26] | 77.76 | 83.15 | 80.45 | 0.610 |
| K-Mer - Support Vector Machine[27] | 79.48 | 78.04 | 76.61 | 0.561 |

The significant values for every metric are indicated by highlighted values.

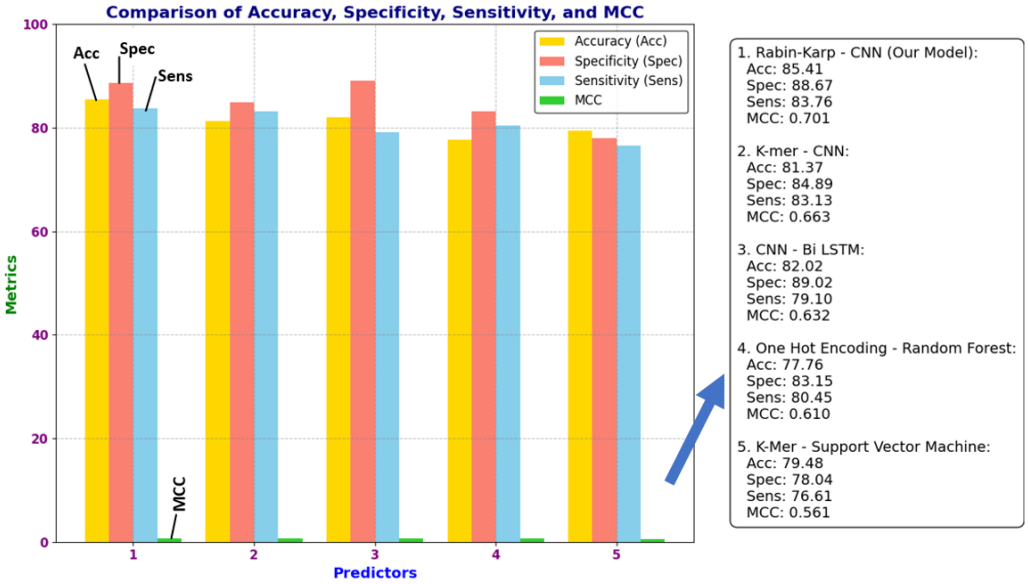



Fig 9. Performance comparison of models.


Fig 10. Output window for Transcription (DNA to RNA), Translation (RNA to Protein), and mutation detection in p53 DNA sequence.

## VIII. Implementation and Results

TABLE III. RESULT: DETECTION OF POSSIBLE MUTATIONS IN THE P53 PROTEIN.

| S. No | Exon | Codon | Expression of Wild Allele | Mutation | Nucleotide Change | Amino Acid Change |
|---|---|---|---|---|---|---|
| 1 | 2--3--4 | 1 -> 43 | + | Another polypeptide sequence in N terminus | 1 -> 43 | 1 -> 129 |
| 2 | 4 | 33 -> 48 | + | Another polypeptide sequence between AA 33 and 48 | 33 -> 48 | 97 -> 144 |
| 3 | 5 | 143 | + | V143A | GTG -> GCG | Valine (V) -> Alanine (A) |
| 4 | 5 | 143 | - | V143E | GTG -> GAG | Valine (V) -> Glutamic Acid (E) |
| 5 | 5 | 161 | - | A161T | GCC -> ACC | Alanine (A) -> Threonine (T) |
| 6 | 5 | 167 | + | E167R | CAG -> CGG | Glutamic Acid (E) -> Arginine (R) |
| 7 | 5 | 168 | + | H168R | CAC -> CGC | Histidine (H) -> Arginine (R) |
| 8 | 5 | 175 | - | R175H | CGC -> CAC | Arginine (R) -> Histidine (H) |
| 9 | 5 | 175 | - | R175G | CGC -> GGC | Arginine (R) -> Glycine (G) |
| 10 | 5 | 177-179 | + | Deletion | Deletion of CCC CAC CAT | Deletion of PHH |
| 11 | 5 | 179 | + | H179Y | CAT ->TAT | Histidine (H) -> Tyrosine (Y) |
| 12 | 6 | 194 | + | L194P | CTT - > CCT | Leucine (L) -> Proline (P) |
| 13 | 6 | 196 | + | R -> STOP | CGA -> TGA | R -> STOP |
| 14 | 6 | 205 | + | Y205D | TAT -> GAT | Tyrosine (Y) -> Aspartic Acid (D) |
| 15 | 6 | 220 | - | Y220C | TAT -> TGT | Tyrosine (Y) -> Cysteine (C) |
| 16 | 7 | 245 | + | G245D | GGC -> GAC | Glycine (G) -> Aspartic Acid (D) |
| 17 | 7 | 245 | + | G245C | GGC -> TGC | Glycine (G) -> Cysteine (C) |
| 18 | 7 | 248 | - | R248Q | CGG -> CAG | Arginine (R) -> Glutamine (Q) |
| 19 | 7 | 248 | - | R248W | CGG - > TGG | Arginine (R) -> Tryptophan (W) |
| 20 | 7 | 248 | + | R248Q | CGG -> CAG | Arginine (R) -> Glutamine (Q) |
| 21 | 7 | 255 | + | I255F | ATC -> TTC | Isoleucine (I) -> Phenylalanine (F) |
| 22 | 7 | 257 | + | Frameshift mutation | Deletion CTG -> TG | Frameshift changing AA sequence [stop codon 344] |
| 23 | 7 | 259 | + | Frameshift mutation | Insertion GAC -> GTAC | Frameshift changing AA sequence [stop codon 263] |
| 24 | 8 | 263 | + | Frameshift mutation | Insertion of 5b: AA [GGTAA] T | Frameshift changing AA sequence [stop codon 344] |
| 25 | 8 | 266 | - | G266V | GGA -> GTA | Glycine -> Valine (V) |
| 26 | 8 | 272 | - | V272M | GTG -> ATG | Valine (V) -> Methionine (M) |
| 27 | 8 | 272 | - | V272L | GTG -> TTG | Valine (V) -> Leucine (L) |

| | | | | | | |
|---|---|---|---|---|---|---|
| 28 | 8 | 273 | - | R273C | CGT -> TGT | Arginine (R) -> Cysteine (C) |
| 29 | 8 | 273 | + | R273H | CGT -> CAT | Arginine (R) -> Histidine (H) |
| 30 | 8 | 273 | - | R273H | CGT -> CAT | Arginine (R) -> Histidine (H) |
| 31 | 8 | 274 | - | V274L | GTT -> CTT | Valine (V) -> Leucine (L) |
| 32 | 8 | 280 | + | Frameshift mutation | Deletion AGA -> GA | Frameshift changing AA sequence [stop codon 344] |
| 33 | 8 | 301 | + | Frameshift mutation | Deletion CCA ->CA | Frameshift changing AA sequence [stop codon 344] |
| 34 | 9 | 306 | + | R -> STOP | CGA -> TGA | R -> STOP |

| Sequence Logo | Genetic Alteration and Translational Impact |
|---|---|
|  | **Exon 5, Codon 143 Mutation** This mutation alters an essential amino acid in the DNA-binding domain, potentially disrupting TP53's ability to regulate target genes. |
|  | **Exon 5, Codon 161 and 167 Mutations** These mutations occur within the structured region of the TP53 protein, which may impact its stability or interaction with cofactors. |
|  | **Exon 5, Codon Deletion 177-179** The deletion of codons in this region results in the loss of amino acids critical for proper protein folding and function, possibly leading to a dysfunctional TP53 protein. |
|  | **Exon 9, Codon 306 Stop Mutation** This is a premature stop codon mutation, leading to a truncated TP53 protein that lacks crucial functional domains significantly impairing its tumor-suppressor activity. |
|   | **Exon 7, Codon 257 Frameshift Mutation** A frameshift mutation alters the reading frame, producing an aberrant and likely non-functional protein, which may contribute to oncogenesis. |

Fig 11. Representation of genetic disruption and premature termination.

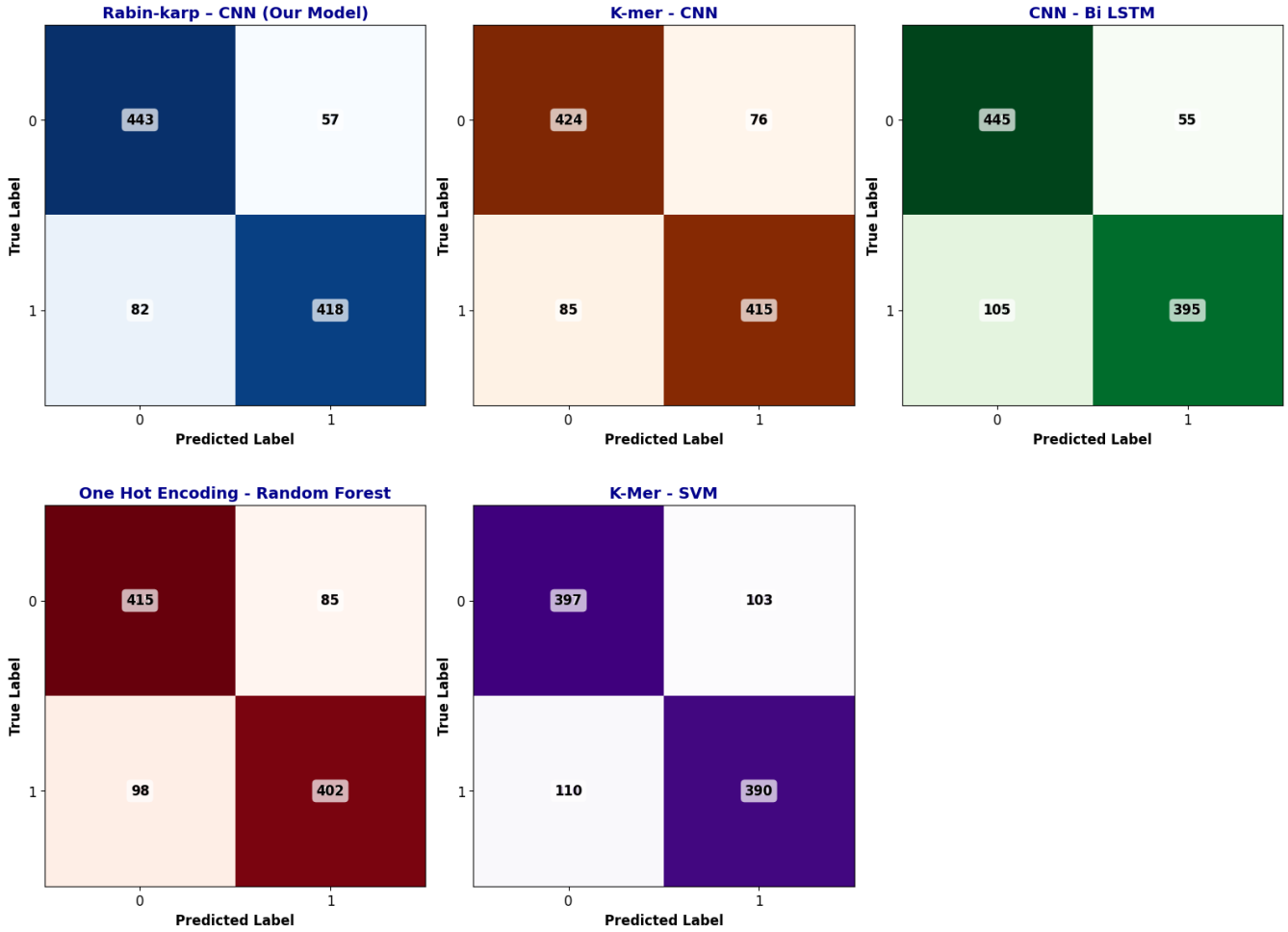



Fig 12. Confusion matrix representation for each model. Superior performance of Rabin-Karp with CNN compared to other models.

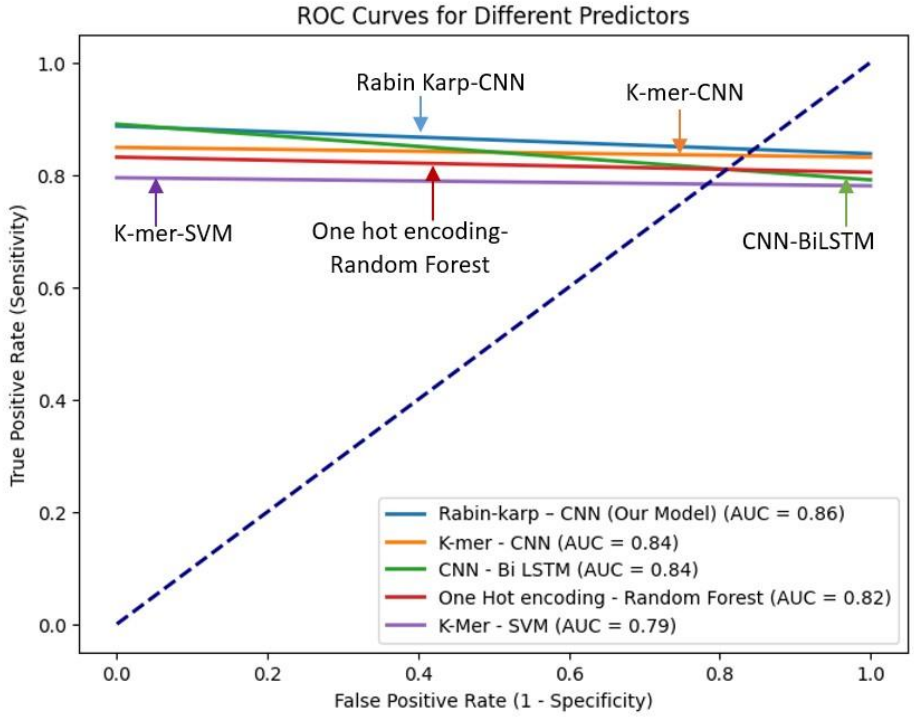
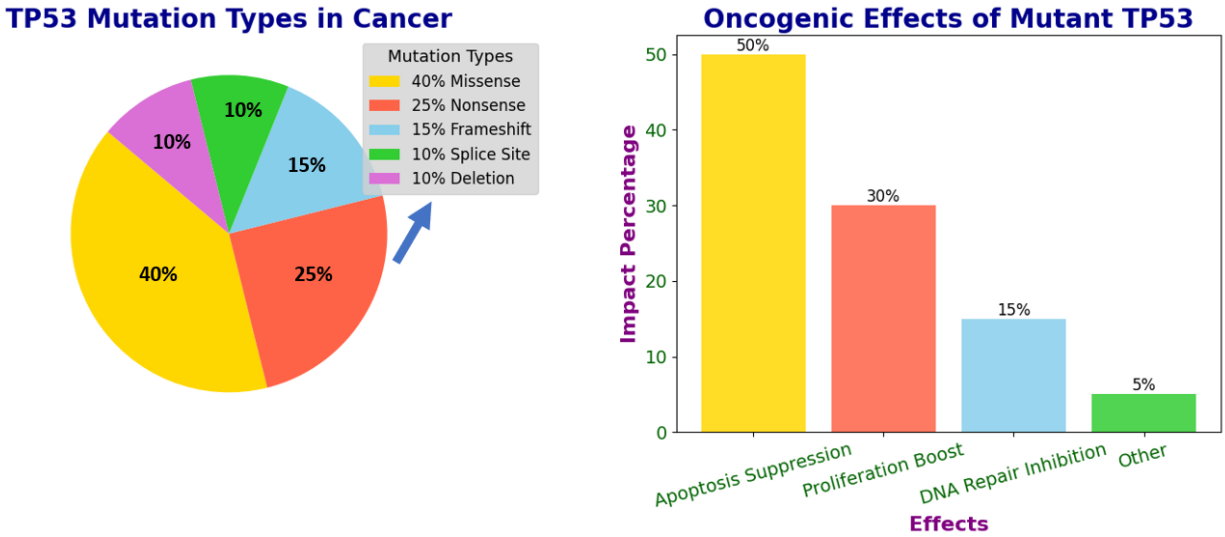



Fig 13. ROC Curve comparison of the models.




(a) Various types of cancer in mutant TP53. (b) Causes and effects percentage of mutant TP53

Fig 14. Analysis of mutant TP53 in cancer types and its effects

The result in Figure 12 is a series of observable confusion matrices related to the prediction performance of the models used in the classification task. Each statistic is associated with a specific prediction, including "Rabin-Karp – CNN (Our Model)," "K-mer – CNN", "CNN - Bi LSTM", "One Hot encoding - Random Forest", and "K-mer - SVM". These confusion matrices help to evaluate true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) for each model, and offering a detailed perspective on their performance. Each image is a 2x2 grid. The cells in the top left (TN) represent correctly detected negative cases. The cells (TP) in the lower right corner indicate positively identified individuals. The cells in the upper right corner (FP) show the best classification of negative cases. The cells in the bottom left (FN) contain positive cases that are incorrectly classified as negative.

The predictor "Rabin-Karp – CNN (our model)" has the highest accuracy and the best balance of sensitivity and specificity, which shows strong performance in correctly detecting positive and negative information. The "K-mer - CNN" and "CNN - Bi LSTM" models also exhibit competitive results, with slightly higher TP and TN. "One Hot encoding - Random Forest" and "K-Mer - SVM" show weaker performance, with increasing values of FP and FN, which can negatively affect both their accuracy and reliability. The system compares the performance of prediction models for mutation data analysis based on four metrics: accuracy, sensitivity, specificity, and Matthew's correlation coefficient (MCC).

Among the evaluated models "Rabin-Karp - CNN (our model)" performs better than the "K Mer - CNN", "CNN -". BLSTM", "one-hot encoding - Random Forests", and "K-Mer - SVM" as shown in Figure 9. Bar height for each model represents the percentage of each metric. "Accuracy," "sensitivity," and "specificity" generally show high values (greater than 80%) in all models, indicating strong predictive power. Among the models, the "Rabin-Karp - CNN" exhibits a competitive performance in all metrics, indicating its effectiveness for mutation analysis. For other metrics (sensitivity, specificity), the differences between models appear to be small, suggesting similar performance in these aspects.

The table III presents various TP53 mutations, detailing their exon and codon locations, expression status, mutation types, nucleotide changes, and corresponding amino acid alterations. Several mutations occur within exon 5, including V143A (Valine to Alanine) and R175H (Arginine to Histidine), among others. Some mutations result in complete amino acid deletions, such as codon 177-179 deletion, while others introduce frameshift mutations, altering the protein's reading frame and potentially leading to premature stop codons. Exon 7 harbors multiple mutations, including R248Q and R248W, both affecting arginine at codon 248. Additionally, exon 8 exhibits R273H and R273C mutations, which substitute arginine with histidine or cysteine, respectively. Frameshift mutations are also observed in exons 7, 8, and 9, often leading to disrupted protein function. The presence of stop codon mutations in exons 6 and 9 further highlights the impact of these alterations on TP53 protein integrity and its role in tumor suppression.

Figure 13 shows the ROC curve comparison for our model with others. This comparison highlights the reliability and effectiveness of deep learning-based models, especially discrete CNNs, in detecting and predicting mutations.

The first figure 14 (a) pie chart titled "TP53 Mutation Types in Cancer," which breaks down into the different mutation types. The larger block containing 40% represents missense mutations. This is followed by 25% of nonsense mutations, 15% of frameshift mutations, both deletion and splice site mutations, each contributing 10% to the total. The second figure 14 (b) bar graph titled "Oncogenic effects of mutant TP53", showing the effect percentages of the different oncogenic effects. The most prominent effect is the inhibition of apoptosis, and the percentage of results is the highest. This is followed by diffusion enhancement, which has a moderating effect. Inhibition of DNA repair exhibits a minor effect, with the other labeled side producing the lowest percent effect.

TABLE IV. INTENSITY OF THE MUTATION IS CALCULATED FOR THIS MUTANT SEQUENCE

| Sequence Number | Mutant TP53 Sequences | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | DNA Position | | | | | | | | | | | | | | | | | | | |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| *Sequence 1* | *T* | *A* | *G* | *T* | *T* | *G* | *T* | *A* | *C* | *A* | *T* | *T* | *A* | *T* | *T* | *C* | *A* | *T* | *G* | *T* |
| *Sequence 2* | *G* | *G* | *A* | *G* | *C* | *A* | *C* | *T* | *C* | *T* | *G* | *A* | *A* | *G* | *G* | *A* | *C* | *A* | *A* | *C* |
| *Sequence 3* | *C* | *T* | *G* | *T* | *G* | *T* | *C* | *T* | *T* | *G* | *C* | *A* | *G* | *C* | *A* | *G* | *C* | *A* | *G* | *T* |
| *Sequence 4* | *C* | *G* | *G* | *C* | *A* | *T* | *T* | *C* | *G* | *C* | *A* | *T* | *A* | *A* | *G* | *C* | *T* | *C* | *A* | *G* |
| *Sequence 5* | *A* | *T* | *A* | *T* | *G* | *G* | *C* | *T* | *G* | *C* | *A* | *T* | *C* | *C* | *A* | *G* | *T* | *G* | *C* | *G* |

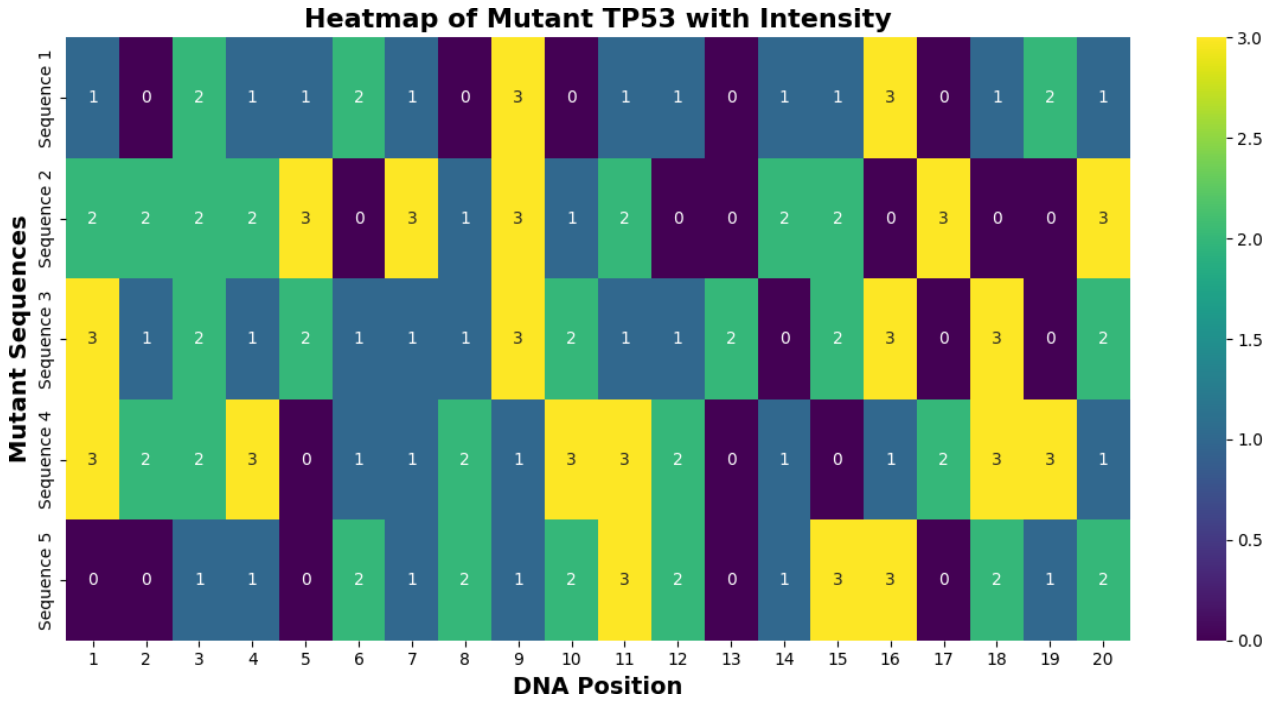



Fig 15. Heatmap of mutation with its intensity.

Table IV is the connectivity for Figure 15. It represents a heat map of the mutated TP53 DNA sequence, showing the variation in DNA positions among the five mutated sequences. Sequences correspond to individual sequence mutations (numbered sequence 1 to sequence 5), while columns represent specific DNA regions, numbered 1 to 20. The color intensity of each cell indicates the degree of mutation is the sensitivity or frequency of occurrence at that locus, where the color bar on right indicates mutation values (from 0 to 3).

For example, dark purple cells (value 0) indicate no mutations in that area, while bright red cells (value 3) represent the most severe mutations. This heat map visually reveals patterns and differences in mutations across sequences, identifying conserved and highly variable regions of DNA for example 9 and 16 positions significant sequence change is evident, as indicated by fluorescent cells, indicating potential hotspots for mutations. Positions of the same color, such as position 8, indicative of overlapping elements so the stability is contradicted. Overall, this heat map is a tool to identify and analyze mutation distribution and intensity in TP53 mutated sequences.

The specificity scores of TP53 frameshift mutations at various genomic regions are exposed in Figure 16.

The y-axis displays the specificity scores ranging from 0.8 to 0.9, while the x-axis displays the chromosomal sites where the mutations occur. The graphic representation of each mutation as a labeled point, where the labels denote the particular nucleotide changes (e.g., "AAT->G," "CCG->T"). It is simple to match the genomic site with the particular mutation and its associated specificity score is marked as a red text annotation, which offer comprehensive information about the type of mutation at each place. The specificity ratings exhibit slight fluctuations across the various mutation locations, although they often stay high (over 0.8).

Several TP53 mutations and the accompanying specificity scores are displayed in a horizontal bar chart in Figure 17. A particular mutation is represented by each bar, and the y-axis labels the specifics of the mutation (e.g., "G->A @ 175," "C->T @ 213). From the range of 0.70 to 0.90, the specificity scores are represented on the x-axis. It shows the precise specificity score in bold type next to each bar. For the mutation "A->In @ 420," the specificity score is 0.90, which is the highest; for "G->Stop @ 245," it is 0.70, which is the lowest. The variation in specificity scores between the different TP53 mutation types and their genomic sites are displayed in this chart.

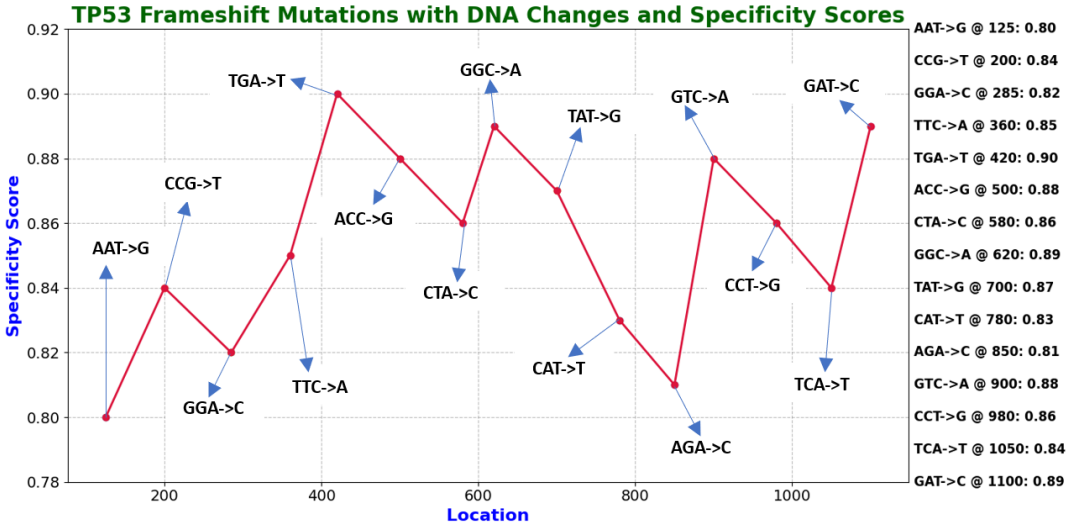



Fig 16. Detection of frameshift mutation in the p53 DNA sequence with locations.

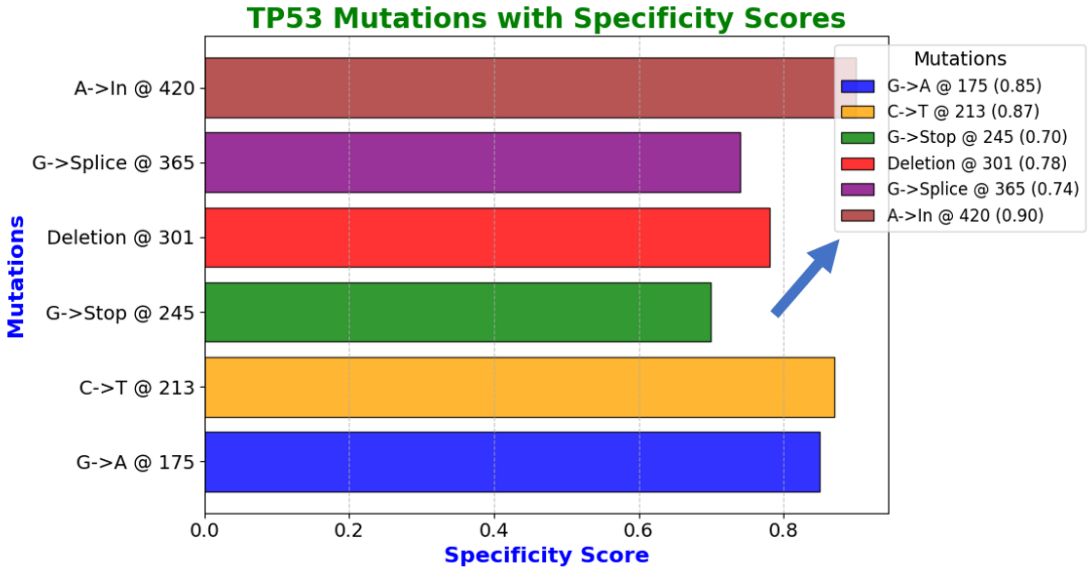



Fig 17. Possible mutations in the p53 DNA sequence.

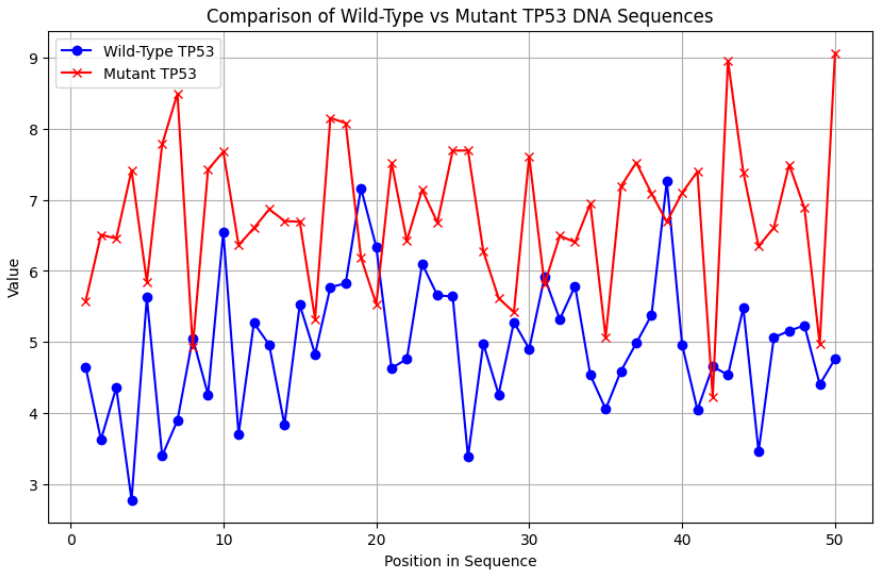



Fig 18. Distinct mutation patterns across DNA sequence positions.

```
Wild-Type DNA: ATGGAGGAGCCGCAGTCAGATCCTAGCGTCGAGAAGTCTGTTATGTG
Wild-Type Protein: MEEPQSDPSVEKSVM
Mutant DNA (substitution): ATGGAGGAGCCGCAGTCATATCCTAGCGTCGAGAAGTCTGTTATGTG
Mutant Protein: MEEPQSYPSVEKSVM
```

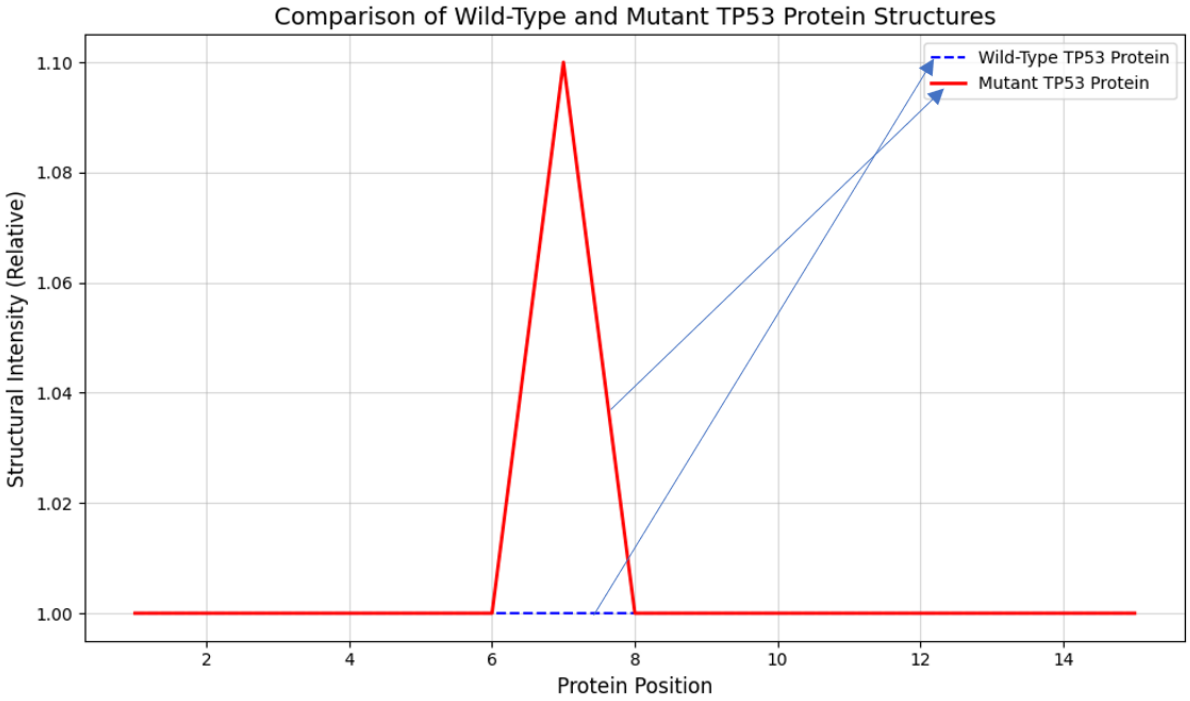



Fig 19. Structural intensity changes in the DNA sequence.

Figure 18 represents the distinct patterns or trends in how the lines behave over positions, it could suggest specific regions where mutations significantly affect function or expression levels. The values for the mutant sequence (red line) tend to be higher than those for the wild-type sequence (blue line). In figure 19 it displays the comparative analysis of wild-type TP53 protein and mutant TP53 protein structural energies (relative) at the protein sequence level. In wild type DNA it is GAT at 19th position but in mutant DNA it is TAT. In wild type protein it is D at 7th position but in mutant protein it is Y. Here the type of mutation is substitution. The blue dashed line represents the wild-type TP53 protein structure, carrying a moving energy value on a constant of 1.00 at all positions, indicating stability and homogeneous structure with no differences. On the contrary, the red line is stable exists for the mutant TP53 protein, showing a significant increase of 7 positions with a magnitude of 1.10. This strong divergence suggests that a mutation has occurred at this particular site, possibly altering the structural or functional integrity of the protein. This peak of the chart best highlights the effect of mutations by contrasting the stable wild type protein with the fluctuating protein.

## IX. Conclusion

The mutation in the p53 protein located in chromosome 17 is the core reason for causing cancer in more than 60% of cases [33]. Since every human comprises of billions of DNA sequences it is hard to mark the exact location and trace where the mutation happens. The p53 protein consists of 393 codon lengths. That is 393 multiples by 3 which is 1179 individual DNA sequences (like ACTGTGTAA...). The list of possible mutations can be detected by marking the change in amino acid with the codon number [34]. This will help us to treat individual cancer patients with specific care and to give suitable tailored medicine.

By detecting the exact mutation in the p53 DNA sequencing targeted therapy can be achieved. According to the mutation that happens in DNA sequence insertion, deletion, or frameshift changes in the codon are identified by the Rabin-Karp algorithm and CNN [35] the specified amino acid changes can be analyzed. Common drugs, chemotherapy, and radiation therapy are not suited for all patients. Some patients may be more sensitive to drugs. Some patients may not respond to the common drugs. Identifying and validating the key mutant gene that causes cancer will help to develop personalized medicine [36]. Personalized medicine will suit the individual patient and rehabilitate cancer with minimal side effects. Advanced research in this DNA sequence mutant-driven area will provide more effective and efficient medicines to positively cure cancer patients.

## References

[1] A. Strasser and D. L. Vaux, “Cell Death in the Origin and Treatment of Cancer,” Mol. Cell, vol. 78, no. 6, pp. 1045–1054, 2020, doi: 10.1016/j.molcel.2020.05.014.

[2] J. Jin et al., “Identification of genetic mutations in cancer: Challenge and opportunity in the new era of targeted therapy,” Front. Oncol., vol. 9, no. MAR, pp. 1–7, 2019, doi: 10.3389/fonc.2019.00263.

[3] N. N. Danial and S. J. Korsmeyer, “Cell Death: Critical Control Points,” Cell, vol. 116, no. 2, pp. 205–219, 2004, doi: 10.1016/S0092-8674(04)00046-7.

[4] E. P. Tan, F. E. Duncan, and C. Slawson, “The sweet side of the cell cycle,” Biochem. Soc. Trans., vol. 45, no. 2, pp. 313–322, 2017, doi: 10.1042/BST20160145.

[5] M. Osmani, M. Tuna, and I. R. Elezaj, “Characterization of nuclear DNA content and chromosome numbers of Tulipa luanica Millaku, T.

kosovarica Kit Tan, Shuka & Krasniqi and T. albanica Kit Tan & Shuka," Acta Agric. Slov., vol. 119, no. 2, 2023, doi: 10.14720/aas.2023.119.2.13280.
[6] S. Benchikh et al., "Chromosome Abnormalities Related to Reproductive and Sexual Development Disorders: A 5-Year Retrospective Study," Biomed Res. Int., vol. 2021, 2021, doi: 10.1155/2021/8893467.
[7] S. Bussolaro et al., "The accuracy of prenatal cell-free DNA screening for sex chromosome abnormalities: A systematic review and meta-analysis," American Journal of Obstetrics and Gynecology MFM, vol. 5, no. 3. 2023, doi: 10.1016/j.ajogmf.2022.100844.
[8] A. Busia et al., "A deep learning approach to pattern recognition for short DNA sequences," bioRxiv, p. 353474, 2019, [Online]. Available: https://www.biorxiv.org/content/10.1101/353474v3%0Ahttps://www.biorxiv.org/content/10.1101/353474v3.abstract.
[9] A. Travers and G. Muskhelishvili, "DNA structure and function," FEBS J., vol. 282, no. 12, pp. 2279–2295, 2015, doi: 10.1111/febs.13307.
[10] F. Mantovani, L. Collavin, and G. Del Sal, "Mutant p53 as a guardian of the cancer cell," Cell Death Differ., vol. 26, no. 2, pp. 199–212, 2019, doi: 10.1038/s41418-018-0246-9.
[11] H. Wang, M. Guo, H. Wei, and Y. Chen, "Targeting p53 pathways: mechanisms, structures, and advances in therapy," Signal Transduction and Targeted Therapy, vol. 8, no. 1. 2023, doi: 10.1038/s41392-023-01347-1.
[12] O. Hassin and M. Oren, "Drugging p53 in cancer: one protein, many targets," Nat. Rev. Drug Discov., vol. 22, no. 2, pp. 127–144, 2023, doi: 10.1038/s41573-022-00571-8.
[13] M. C. Kennedy and S. W. Lowe, "Mutant p53: it's not all one and the same," Cell Death Differ., vol. 29, no. 5, pp. 983–987, 2022, doi: 10.1038/s41418-022-00989-y.
[14] C. A. Garcia, A. Ahmadian, B. Gharizadeh, J. Lundeberg, M. Ronaghi, and P. Nyrén, "Mutation detection by pyrosequencing: Sequencing of exons 5-8 of the p53 tumor suppressor gene," Gene, vol. 253, no. 2, pp. 249–257, 2000, doi: 10.1016/S0378-1119(00)00257-2.
[15] C. Marvalim, A. Datta, and S. C. Lee, "Role of p53 in breast cancer progression: An insight into p53 targeted therapy," Theranostics, vol. 13, no. 4, pp. 1421–1442, 2023, doi: 10.7150/thno.81847.
[16] S. Elmore, "Apoptosis: A Review of Programmed Cell Death," Toxicol. Pathol., vol. 35, no. 4, pp. 495–516, 2007, doi: 10.1080/01926230701320337.
[17] R. S. Y. Wong, "Apoptosis in cancer: From pathogenesis to treatment," J. Exp. Clin. Cancer Res., vol. 30, no. 1, pp. 1–14, 2011, doi: 10.1186/1756-9966-30-87.
[18] K. Duvvuri, P. N. Reddy, H. Kanisettypalli, D. Radha, and T. V. Nidhin Prabhakar, "Comparative Analysis of Pattern Matching Algorithms Using DNA Sequences," 2022, doi: 10.1109/MysuruCon55714.2022.9972412.
[19] C. Xi et al., "AI-BRAF V600E: A deep convolutional neural network for BRAF V600E mutation status prediction of thyroid nodules using ultrasound images," View, no. December 2022, p. 20220057, 2023, doi: 10.1002/viw.20220057.
[20] Siyu Jia, and Ying Tian, "Face Detection Based on Improved Multi-task Cascaded Convolutional Neural Networks," IAENG International Journal of Computer Science, vol. 51, no. 2, pp67-74, 2024.
[21] M. Uddin, M. K. Islam, M. R. Hassan, F. Jahan, and J. H. Baek, "A fast and efficient algorithm for DNA sequence similarity identification," Complex Intell. Syst., vol. 9, no. 2, pp. 1265–1280, 2022, doi: 10.1007/s40747-022-00846-y.
[22] L. U. Akhmetzianova, T. M. Davletkulov, I. M. Gubaidullin, and A. R. Islamgulov, "Parallel implementation of the primer search algorithm for loop-mediated isothermal amplification," in Journal of Physics: Conference Series, 2021, vol. 2131, no. 2, doi: 10.1088/1742-6596/2131/2/022004.
[23] Yong Cheng, Huawei Tao, Xiangying Lu, Zhijun Xu, Hongliang Fu, and Chunhua Zhu, "Prediction Method of Axial Compression Capacity of CCFST Columns Based on Deep Learning," IAENG International Journal of Computer Science, vol. 51, no. 3, pp243-251, 2024
[24] K. Aygen, U. Eser, and I. B. Celik, "Determining Positive Cancer Rescue Mutations in p53 Based Cancers by using Artificial Intelligence," arXiv, vol. 53, 2017.
[25] P. Dataset, "NCBI," https://www.ncbi.nlm.nih.gov/gene/?term=p53.
[26] H. Gunasekaran, K. Ramalakshmi, A. Rex Macedo Arokiaraj, S. D. Kanmani, C. Venkatesan, and C. S. G. Dhas, "Analysis of DNA Sequence Classification Using CNN and Hybrid Models," Comput. Math. Methods Med., vol. 2021, 2021, doi: 10.1155/2021/1835056.
[27] Yong Cheng, Huawei Tao, Xiangying Lu, Zhijun Xu, Hongliang Fu, and Chunhua Zhu, "Prediction Method of Axial Compression Capacity of CCFST Columns Based on Deep Learning," IAENG International Journal of Computer Science, vol. 51, no. 3, pp243-251, 2024
[28] U. N. Wisesty et al., "Join Classifier of Type and Index Mutation on Lung Cancer DNA Using Sequential Labeling Model," IEEE Access, vol. 10, 2022, doi: 10.1109/ACCESS.2022.3142925.
[29] L. Muflikhah, M. A. Rahman, and A. W. Widodo, "Profiling DNA sequence of SARS-Cov-2 virus using machine learning algorithm," Bull. Electr. Eng. Informatics, vol. 11, no. 2, 2022, doi: 10.11591/eei.v11i2.3487.
[30] F. L. Heredia, A. Roche-Lima, and E. I. Parés-Matos, "A novel artificial intelligence-based approach for identification of deoxynucleotide aptamers," PLoS Comput. Biol., vol. 17, no. 8, 2021, doi: 10.1371/journal.pcbi.1009247.
[31] [S. G. Jacob, B. Bennet, and M. M. Ali Sulaiman, "Investigating the Role of PTEN and P53 in Autism: Design of A Mutant Information Prediction System (MIPS)," 2023, doi: 10.1109/ICCAE56788.2023.10111310.
[32] K. Hernández-Vidales, E. Guevara, V. Olivares-Illana, and F. J. González, "Characterization of wild-type and mutant p53 protein by Raman spectroscopy and multivariate methods," J. Raman Spectrosc., vol. 50, no. 10, 2019, doi: 10.1002/jrs.5655.
[33] Khaoula Slime, Abderrahim Maizate, Larbi Hassouni, and Najat Mouine, "Toward a Model to Predict Cardiovascular Disease Risk Using a Machine Learning Approach," IAENG International Journal of Computer Science, vol. 51, no. 5, pp519-527, 2024
[34] G. Zhu et al., "Mutant p53 in Cancer Progression and Targeted Therapies," Front. Oncol., vol. 10, no. November, pp. 1–9, 2020, doi: 10.3389/fonc.2020.595187.
[35] Marizuana Mat Daud, Zulaikha Kadim, and Hon Hock Woon, "Detection of Oil Palm Tree and Loose Fruitlets for Fresh Fruit Bunch's Ready-to-Harvest Prediction via Deep Learning Approach," IAENG International Journal of Computer Science, vol. 50, no.4, pp1183-1193, 2023
[36] G. Alliance, "Understanding Genetics: A New York, Mid- Atlantic Guide for Patients and Health Professionals," Genet. Alliance; New York - Mid-Atlantic Consort. Genet. Newborn Screen. Serv., no. Appendix E Inheritance Patterns, p. 105, 2009.

**Aruna Arumugam. G** received her B.Sc. degree in Computer Science at Sarah Tucker College, Palayamkottai, Tirunelveli in 2008. She then obtained her MCA from Manonmaniam Sundaranar University, Tirunelveli in 2011, followed by an M.Phil. from St. Peter's University, Chennai in 2014. With 9 years of teaching experience, she is currently pursuing her Ph.D. in the field of machine learning at B.S. Abdur Rahman Crescent Institute of Science and Technology, Chennai. Her areas of interest include artificial intelligence, deep learning, machine learning, computational biology, computer networks, and programming languages.

**Dr. Mohamed Divan Masood** is an Assistant Professor in the Department of Computer Applications at the B.S. Abdur Rahman Crescent Institute of Science & Technology, Chennai a position he has held since February 2019. Dr. Masood earned his Ph.D. in Computer Science and Engineering from Anna University, Chennai, in 2018. He holds an M.Tech in Information Technology from B.S. Abdur Rahman Crescent Institute of Science & Technology, obtained in 2013, and a B.Tech in Information Technology from Anna University, Tirunelveli, completed in 2011. His research interests encompass a broad spectrum of cutting-edge areas, including Computational Biology, Artificial Intelligence, Machine Learning, Deep Learning, Big Data, and Social Network Analysis.