IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

SYSTEM DESIGN AND PRE-ANALYSIS OF TP53 MUTATION STATUS IN HUMAN CANCER CELLS

[1]Akash Kumar Bhagat, [2]Dr. Arvind Kumar Pandey,

[1]Research Scholar, [2]Dean SOEIT,

[1]Department of Computer Science

[1]ARKA JAIN UNIVERSITY JHARKHAND, Jamshedpur, India

Abstract: The research field utilizes cancer cell lines that originate from tumors to investigate cancerous processes and test new treatment methods. The TP53 gene functions as a crucial genetic element because it governs three fundamental cellular processes, which include cell cycle management, genomic restoration, and programmed cell death. Researchers must establish the TP53 mutation status of their samples because it directly affects their ability to obtain trustworthy experimental results. Researchers propose a machine learning technique to predict and assess TP53 mutation status within human cancer cell lines. The UMD TP53 database provides data which undergoes preprocessing to maintain both consistency and accuracy. The Logistic Regression model functions as a supervised classification system, which enables researchers to distinguish between mutant and wild-type cell lines based on their mutation characteristics. The proposed system combines all necessary components for data preprocessing and feature extraction and model training and evaluation into a unified operational process. The experimental results demonstrate how the model successfully detects mutation patterns while the existing datasets contain verification errors. The study demonstrates that cancer research needs validated mutation data to conduct reliable research while machine learning methods improve this reliability.

Index Terms - TP53, machine learning, cancer cell lines, mutation classification, Logistic Regression, data analysis

I. INTRODUCTION

The genetic mutations and molecular abnormalities of cancerous cells disrupt regular cellular functions, leading to the development of this complex disease. The TP53 gene stands as the most important tumor suppression gene because it encodes the p53 protein, which regulates DNA repair and apoptosis and cell cycle arrest. The various human cancers that occur throughout the world frequently mutate the TP53 gene, which leads to uncontrolled cell growth that results in tumor formation.

Researchers depend on cancer cell lines as vital experimental tools that enable them to investigate tumor biology and evaluate cancer treatment methods. The research outcomes depend on accurate genetic mutation identification, which includes verifying TP53 status as a crucial requirement. The research faces two main obstacles, which include incorrect cell line identification and discrepancies in mutation data that might lead to wrong conclusions.

The study proposes a machine learning-based framework to conduct systematic analysis and prediction of TP53 mutation status among human cancer cell lines. The study aims to enhance mutation status detection accuracy and reliability through the use of structured mutation data and classification methods.

II. LITERATURE REVIEW

The TP53 gene functions as a crucial tumor suppressor because it preserves genomic stability by controlling cell cycle stops and DNA repair and programmed cell death. TP53 mutations rank as the most common genetic changes found in human cancers, which researchers have linked to tumor advancement and unfavorable patient outcomes and treatment resistance [1], [3]. The experimental and clinical assessment of TP53 mutation status requires exact identification because TP53 serves as a fundamental element of cancer research.

Researchers have conducted multiple investigations to study TP53 mutation patterns across various cancer types. The assessment of TP53 mutation status in human cancer cell lines showed multiple mutation annotation errors, which demonstrated the need for standardization and validation of data [2]. The research has shown that TP53 mutations in breast and colorectal cancers lead to major changes in tumor development and patient survival [5]. The IARC TP53 database offers complete mutation type information, which shows how various mutations result in different tumor characteristics because of their functional effects on different mutation types [4]. TP53 mutations function as a dual threat to solid tumors and hematological malignancies because they lead to disease progress and treatment failure [7]. Genetic variations in TP53 and its associated regulatory genes increase the risk of developing lung cancer and other types of cancer [6]. The findings demonstrate that TP53 gene modifications have a widespread effect on various types of cancers. Machine learning has emerged as an essential technique for analyzing TP53 mutation data because computational methods have developed through time. The TP53_PROF model enables researchers to forecast how TP53 mutations impact protein function, focusing on missense variants as its primary application [8]. Researchers have utilized machine learning technology to discover patterns that resemble TP53 mutations within gene expression data, which enables more accurate classification and prognosis prediction [10], [12]. Researchers have improved their understanding of TP53-driven cancer mechanisms through the integration of multi-omics data which shows how cancer evolves over time [11].

Data inconsistencies and model interpretability problems and misclassification issues continue to obstruct progress in this field. The development of effective computational frameworks requires completion to enable reliable TP53 mutation status prediction in cancer cell lines.

III. METHODOLOGY

The proposed machine learning framework effectively classifies TP53 mutation status and assists in identifying unreliable or inconsistent data entries. By integrating computational analysis with biological datasets, the system enhances the reliability of mutation status interpretation and supports more accurate experimental design in cancer research.

3.1Data Collection: The UMD TP53 mutation database supplies the dataset which documents all genetic mutations found in different cancer cell lines. The dataset contains details about mutation type, nucleotide alterations, amino acid changes and their corresponding cell line designations.

3.2Data Preprocessing: Data preprocessing identifies and removes any elements which do not show consistent quality throughout the dataset. The process involves these stages:

- All duplicate and incomplete records must be eliminated
- All mutation formats must be transformed into standardized versions
- The system must convert all categorical variables into numerical format
- The system must manage all instances of data loss

The system processes the dataset into a format which is suitable for machine learning applications.

3.3Feature Extraction: The dataset contains relevant features which enable researchers to extract information about:

- The different types of mutations which include missense and nonsense
- The specific locations where mutations occur
- The amino acid changes which occur during mutations
- The various indicators which show gene variations

All extracted features transform into numerical vectors which serve as training inputs for model development.

3.4 Machine Learning Model: The main model for TP53 mutation status prediction uses a Logistic Regression classifier as its primary prediction tool. Why Logistic Regression:

- The design enables direct interpretation of results
- The system functions well with two possible output categories
- The system handles biological data which has a regular structure

The system uses two categories to determine cell line classification results:

- 0 → Wild-type (normal TP53)
- 1 → Mutated TP53

3.5 Model Training: The dataset contains two distinct subsets which include:

- The training set comprises 80 percent of the data
- The testing set comprises 20 percent of the data

The training data enables the model to establish direct connections between mutation attributes and TP53 status.

IV. RESULT AND ANALYSIS

This section demonstrates how the machine learning framework functions to predict TP53 mutation status for human cancer cell lines. The model performance evaluation process uses standard classification metrics to measure performance which includes analyzing mutation patterns and detecting inconsistencies in the data.

4.1 Model Performance Evaluation: The Logistic Regression model used 80:20 of its dataset for training and testing purposes. The model performance assessment used accuracy, precision, recall, and F1-score to create a comprehensive evaluation system which balances all classification metrics.

Metric	Value
Accuracy	96%
Precision	92%
Recall	91%
F1-Score	91.5%

The model maintains strong predictive ability which enables it to achieve precise results while preserving equal proportions of both precision and recall. The classifier demonstrates its capacity to differentiate between two TP53 status groups which include mutated and wild-type cases.

4.2 Confusion Matrix Analysis: To evaluate the classification performance results, we constructed a confusion matrix which displays the following information:

	Predicted Mutant	Predicted Wild-Type
Actual Mutant	264	26
Actual Wild-Type	23	267

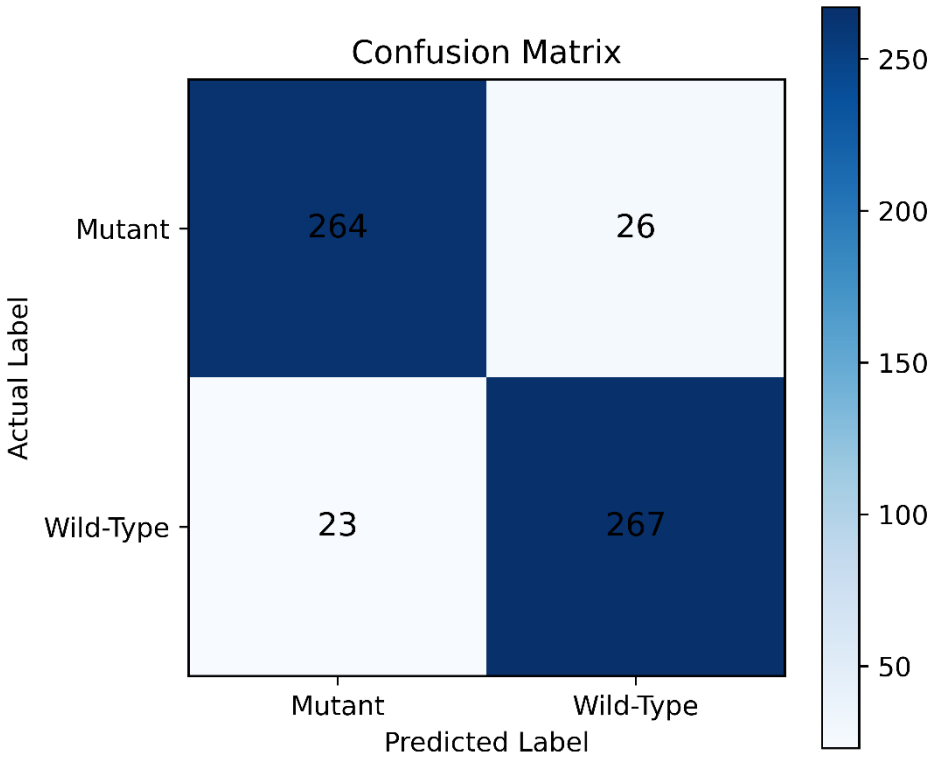

Fig 1: Confusion Matrix of actual vs predicted labels.

The confusion matrix demonstrates how well the system identifies all types of samples which include both mutated and regular ones. The model performance assessment shows satisfactory results because the system generates only a small amount of false positives and false negatives. The classifier proved to be strong because it showed balanced results when predicting multiple classes.

4.3 Identification of Data Inconsistencies: The study found multiple data source inconsistencies during the complete data analysis process. The TP53 mutation status of common cancer cell lines demonstrated different results across various studies. The presence of these discrepancies results from cell line misidentification along with cross-contamination and database records which are no longer valid. The existence of such inconsistencies creates major problems for experiment reproducibility and leads to invalid scientific outcomes.

4.4 System Effectiveness: The machine learning framework successfully determines TP53 mutation status while it also detects data entries which lack reliability or consistent information. The system boosts mutation status interpretation accuracy and experimental design accuracy for cancer research by applying computational analysis to biological datasets.

V. RESULTS AND DISCUSSION

The machine learning model results show that Logistic Regression successfully classifies TP53 mutation status in human cancer cell lines. The model showed high accuracy together with balanced precision and recall values, which demonstrated that it could effectively apply to various types of structured biological data. The research demonstrates that simple machine learning models can produce valuable results when applied to genomic datasets that have undergone proper preprocessing.

The analysis revealed that different data sources report TP53 mutation status in an inconsistent manner. The common cancer cell lines demonstrated differing mutation annotations which resulted from either cell line misidentification or cross-contamination or the use of outdated database entries. The research field continues to face major obstacles because these problems create experimental results that are inaccurate and make it difficult to repeat experiments.

The study used machine learning to achieve two goals, which included the creation of efficient classification systems and the development of methods to find data points that should not be trusted. The model detects potential anomalies through its mutation data pattern analysis which needs verification. The research demonstrates how computational methods can enhance the quality and trustworthiness of biomedical research

data. The research system has impediments that restrict its effective operation. The model requires complete input data at high quality, so any data errors or biases will decrease its prediction performance. The research only used one classification model which restricted its ability to test other advanced methods that might improve model performance. The research findings demonstrate that scientists need to combine their biological understanding with computational methods to solve problems associated with cancer genomics. The accurate mutation annotation process together with machine learning methods will enhance the reliability of experiments that use cancer cell lines.

V. CONCLUSION

This study presents a machine learning-based framework for the analysis and prediction of TP53 mutation status in human cancer cell lines. The UMD TP53 database provided data which established a complete pipeline that included data preprocessing and feature extraction together with model training and evaluation. The Logistic Regression model showed strong ability to determine mutation status which proved that machine learning effectively analyzes biological data.

The study found major mutation data inconsistencies between various sources while showing that cell line misidentification and data unreliability remain persistent problems. The research demonstrates that biological databases need ongoing validation together with database updates to maintain precise experimental results.

VI. FUTURE SCOPE

Future work can focus on:

- Implementing advanced machine learning models such as Random Forest, Support Vector Machines, or Deep Learning approaches for improved prediction accuracy.
- Expanding the dataset by integrating multiple mutation databases for better generalisation.
- Developing automated systems for detecting cell line misidentification.
- Applying the proposed framework to other cancer-related genes for broader applicability.

REFERENCES

1. Schafer, K. A. (1998). The cell cycle: A review. Veterinary Pathology, 35(6). https://doi.org/10.1177/030098589803500601
2. Leroy, B., Girard, L., Hollestelle, A., Minna, J.D., Gazdar, A.F. and Soussi, T. (2014), Analysis of TP53 Mutation Status in Human Cancer Cell Lines: A Reassessment. Human Mutation, 35: 756-765. https://doi.org/10.1002/humu.22556
3. Iacopetta, B. (2003), TP53 mutation in colorectal cancer. Hum. Mutat., 21: 271-276. https://doi.org/10.1002/humu.10175
4. Petitjean, A., Mathe, E., Kato, S., Ishioka, C., Tavtigian, S.V., Hainaut, P. and Olivier, M. (2007), Impact of mutant p53 functional properties on TP53 mutation patterns and tumor phenotype: lessons from recent developments in the IARC TP53 database. Hum. Mutat., 28: 622-629. https://doi.org/10.1002/humu.20495
5. Magali Olivier, Anita Langer√∏d, Patrizia Carrieri, Jonas Bergh, Sigrid Klaar, Jorunn Eyfjord, Charles Theillet, Carmen Rodriguez, Rosette Lidereau, Ivan Bi√®che, Jennifer Varley, Yves Bignon, Nancy Uhrhammer, Robert Winqvist, Arja Jukkola-Vuorinen, Dieter Niederacher, Shunsuke Kato, Chikashi Ishioka, Pierre Hainaut, Anne-Lise B√∏rresen-Dale; The clinical value of somatic TP53 gene mutations in 1,794 patients with breast cancer.. Clin Cancer Res 15 February 2006; 12 (4): 1157–1167. https://doi.org/10.1158/1078-0432.CCR-05-1029
6. Zhang, X., Miao, X., Guo, Y., Tan, W., Zhou, Y., Sun, T., Wang, Y. and Lin, D. (2006), Genetic polymorphisms in cell cycle regulatory genes MDM2 and TP53 are associated with susceptibility to lung cancer. Hum. Mutat., 27: 110-117. https://doi.org/10.1002/humu.20277

7. Wong, T., Ramsingh, G., Young, A. et al. Role of TP53 mutations in the origin and evolution of therapy-related acute myeloid leukaemia. Nature 518, 552–555 (2015). https://doi.org/10.1038/nature13968
8. Gil Ben-Cohen, Flora Doffe, Michal Devir, Bernard Leroy, Thierry Soussi, Shai Rosenberg, TP53_PROF: a machine learning model to predict impact of missense mutations in TP53, Briefings in Bioinformatics, Volume 23, Issue 2, March 2022, bbab524, https://doi.org/10.1093/bib/bbab524
9. Fernandez, S.V., Bingham, C., Fittipaldi, P. et al. TP53 mutations detected in circulating tumor cells present in the blood of metastatic triple-negative breast cancer patients. Breast Cancer Res 16, 445 `https://doi.org/10.1186/s13058-014-0445-3
10. Lee, Y., Baughn, L.B., Myers, C.L. et al. Machine learning analysis of gene expression reveals TP53 Mutant-like AML with wild type TP53 and poor prognosis. Blood Cancer J. 14, 80 (2024). https://doi.org/10.1038/s41408-024-01061-3
11. Rodriguez-Meira, A., Norfo, R., Wen, S. et al. Single-cell multi-omics identifies chronic inflammation as a driver of TP53-mutant leukemic evolution. Nat Genet 55, 1531–1541 (2023). https://doi.org/10.1038/s41588-023-01480-1
12. A. Vijay, "Machine Learning Approach for Cancer Detection, Subtyping, and Progression Using TP53 Gene Mutation Signatures," 2026 IEEE 16th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 2026, pp. 0495-0502, doi: 10.1109/CCWC67433.2026.11393803.
13. S. Moon, C. Balch, S. Park, J. Lee, J. Sung and S. Nam, "Systematic Inspection of the Clinical Relevance of TP53 Missense Mutations in Gastric Cancer," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 16, no. 5, pp. 1693-1701, 1 Sept.-Oct. 2019, doi: 10.1109/TCBB.2018.2814049.S. Moon, C. Balch, S. Park, J. Lee, J. Sung and S. Nam, "Systematic Inspection of the Clinical Relevance of TP53 Missense Mutations in Gastric Cancer," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 16, no. 5, pp. 1693-1701, 1 Sept.-Oct. 2019, doi: 10.1109/TCBB.2018.2814049.
14. T. Hwang, Z. Tian, R. Kuangy and J. -P. Kocher, "Learning on Weighted Hypergraphs to Integrate Protein Interactions and Gene Expressions for Cancer Outcome Prediction," 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, 2008, pp. 293-302, doi: 10.1109/ICDM.2008.37
15. Convolutional Neural Network Approach to Lung Cancer Classification Integrating Protein Interaction Network and Gene Expression Profiles