Decoding TP53 Variants: A Statistical and Computational Approach to Prioritize Pathogenic Mutations in Cancer Biology

Usha Adiga[1], Anil B[2], Alfred J Augustine[3], Sampara Vasishta[4]*

[1]Professor, Department of Biochemistry, Apollo institute of medical sciences & Research Chittoor, India
[2]Tutor, Department of Biochemistry, Apollo institute of medical sciences & Research Chittoor, India
[3]Professor, Dept of Surgery, Apollo institute of medical sciences & Research Chittoor, India
*[4]*Research Associate, ICMR, Department of Biochemistry, Apollo institute of medical sciences & Research Chittoor, India*

*** Corresponding Author**
Sampara Vasishta
Research Associate - 1
Department of Biochemistry
Apollo Institute of Medical Sciences and Research
Murukambattu - 517127, Chittoor
Andhra Pradesh, India
Email: vasishta_s@aimsrchittoor.edu.in

KEYWORDS
TP53,guardian of genomes, cancer,pathogenesis

ABSTRACT

Background:
TP53, often referred to as the "guardian of the genome," is a critical tumor suppressor gene that maintains cellular integrity by regulating the cell cycle, DNA repair, and apoptosis. Mutations in TP53 are among the most frequent alterations in cancer and are associated with tumor progression, therapeutic resistance, and poor prognosis. Given the widespread clinical significance of TP53 variants, understanding their functional impact using computational tools has become an essential step in cancer research.***Aim:*** This study aims to comprehensively analyze TP53 variants using a dataset of genomic alterations, focusing on predictive pathogenicity metrics such as SIFT, PolyPhen, CADD, REVEL, MetaLR, and Mutation Assessor. Additionally, the study identifies trends in predictive scores, examines inter-tool correlations, and prioritizes high-risk variants for further clinical investigation.***Materials and Methods:*** A dataset of 3,830 TP53 variants was analyzed. Predictive tools were employed to assess the functional consequences of these variants. Descriptive statistics, correlation analysis, and prioritization criteria based on high CADD (>20), low REVEL (<0.2), and damaging SIFT (≤0.05) scores were applied. Visualizations, including scatter plots and score distributions, were generated to highlight critical insights.***Results:***Descriptive analysis revealed that a majority of TP53 variants have high CADD scores (>20), indicating significant functional impact. SIFT scores clustered near 0, suggesting that many variants are predicted to be damaging. REVEL scores, however, skewed toward lower values, creating discrepancies with CADD. Correlation analysis demonstrated strong agreement between CADD, MetaLR, and Mutation Assessor scores, while REVEL showed weaker correlation with CADD. Variants with high CADD and low REVEL scores were prioritized, as they may represent novel candidates requiring functional validation.**Conclusion:**
This study highlights the diversity and complexity of TP53 variants in cancer, underscoring the importance of integrating multiple predictive scores to assess their pathogenic potential. Variants with discordant predictions between tools may represent unique targets for further experimental validation. Such analyses are essential for identifying clinically actionable TP53 mutations and advancing precision oncology.

INTRODUCTION

The TP53 gene is one of the most extensively researched tumor suppressor genes in the field of cancer biology. It has been granted the title of "guardian of the genome" due to its critical function in preserving cellular homeostasis.[1] The transcription factor p53 is encoded by TP53 and is activated in response to cellular duress, such as oncogene activation, oxidative stress, and DNA damage.[2] Upon activation, p53 coordinates a sequence of critical cellular processes, such as apoptosis, cell cycle arrest, and DNA repair, to prevent the proliferation of damaged cells. This function is essential for the prevention of tumor development and the preservation of genomic stability.[3]

These essential functions are disrupted by mutations in TP53, resulting in genomic instability, uncontrolled cellular proliferation, and tumorigenesis. Based on The Cancer Genome Atlas (TCGA), nearly 50% of instances of human malignancies show TP53 mutations, which underlines their relevance in oncogenesis. Particularly common in breast, lung, ovarian, and colorectal tumors, these mutations not only start the tumor but also help to explain progression, metastases, and medication resistance.[4,5]

The variety of TP53 mutations introduces even another level of complication. These mutations might be missense, changing the structure of the protein, or they can be nonsense and frameshift mutations producing shortened proteins.[6] These mutations have somewhat different functional effects; some cause total loss of tumor suppressor action while others give gain-of- function features that support malignancy. This variability makes it rather difficult to categorize TP53 mutations as benign or harmful.[7]

Bioinformatics has advanced enough that it is now feasible to more precisely forecast the functional effects of genetic mutations. Computational instruments include SIFT, PolyPhen, CADD, REVEL, MetaLR, and Mutation Assessor offer priceless understanding of the pathogenic potential of variations.[8,9] Considering elements like evolutionary conservation, protein structure, and biochemical features, every tool uses a different method. Nevertheless, the different methods of these instruments can lead to different forecasts, which emphasizes the need of integrated techniques to fairly evaluate variant pathogenicity.[10,11]

This work investigates the functional impact of 3,830 TP53 polymorphisms using prediction scores from a well selected dataset. This work intends to close the gap between computational predictions and biological significance by means of descriptive statistics analysis, correlation analysis across predictive methods, and prioritization of high-risk variations. The results not only improve our knowledge of TP53 mutations in cancer biology but also offer a structure for spotting potential variations that may be used as biomarkers for treatment, diagnosis, or prognosis. This work prepares the path for precision oncology approaches aiming at TP53-driven malignancies by combining computational insights with cancer biology.

TP53 orchestrates cellular responses to stress by regulating apoptosis, cell cycle arrest, and DNA repair. However, the functional landscape of TP53 variants remains highly diverse, encompassing missense, nonsense, and frameshift mutations with varying degrees of pathogenicity. These mutations drive tumor progression, metastasis, and resistance to therapy, with nearly 50% of human cancers harboring TP53 alterations.

The diversity of TP53 mutations complicates their classification as benign or pathogenic. Computational tools have emerged as indispensable resources for interpreting genetic variants, yet the variability in predictions across tools necessitates integrative statistical approaches to resolve discrepancies. This study bridges this gap by systematically analyzing 3,830 TP53 variants, leveraging a robust computational framework to provide actionable insights into their functional roles in cancer biology.

Objectives:

1. Conduct a comprehensive statistical analysis of predictive scores to evaluate their distribution, central tendencies, and variability in assessing TP53 variant pathogenicity.

2. Identify significant patterns and interrelationships among predictive tools through correlation analysis and clustering techniques, highlighting areas of agreement and divergence.
3. Develop a robust prioritization framework to identify high-risk pathogenic variants with potential clinical relevance, facilitating their selection for experimental validation and translational research.

MATERIALS AND METHODS

Dataset

The dataset generated from https://asia.ensembl.org/index.html analyzed in this study contains 3,830 TP53 variants with annotations, including chromosome positions, allele types, and predictive scores from tools like SIFT, PolyPhen, CADD, REVEL, MetaLR, and Mutation Assessor. The dataset contained genomic variant information related to the **TP53 gene**. The table included key features like the variant ID, chromosome position, allele types, global minor allele frequency (MAF), and predictive scores for pathogenicity.Variants were systematically cleaned by removing duplicate headers and ensuring numerical consistency.

Methods

Dataset Overview

The dataset used in this study was sourced from the Ensembl genome browser, containing 3,830 TP53 variants. The dataset included detailed annotations such as variant ID, chromosome position, allele types, global minor allele frequency (MAF), and predictive scores from tools like SIFT, PolyPhen, CADD, REVEL, MetaLR, and Mutation Assessor. Data cleaning involved removing duplicate entries, standardizing column headers, and ensuring numerical consistency across all predictive scores. Missing values were imputed where possible to maximize data utility.

Predictive Tools and Scoring Criteria

1. **SIFT**: Predicts whether amino acid substitutions affect protein function. Low scores (≤0.05) indicate likely damaging variants.
2. **PolyPhen**: Evaluates amino acid changes using evolutionary and structural data. Scores closer to 1 predict higher pathogenicity.
3. **CADD**: Integrates multiple annotations to predict deleteriousness. Scores >20 suggest significant functional impact.
4. **REVEL**: Combines scores from multiple tools to predict variant pathogenicity. Higher scores (>0.5) indicate pathogenic potential.
5. **MetaLR**: Uses machine learning to predict deleterious mutations, with scores closer to 1 indicating higher pathogenicity.
6. **Mutation Assessor**: Evaluates functional impact using evolutionary conservation. Scores closer to 1 reflect high impact.

Data Preprocessing and Management

Data preprocessing was carried out using Python (version 3.9) with the Pandas library for data manipulation and cleaning. Missing data were handled using imputation techniques, while non-standard entries were corrected for numerical consistency. Duplicates were systematically removed to ensure a unique dataset. The cleaned dataset was then validated for accuracy by cross-referencing with the original Ensembl data source.

Statistical Analysis

Descriptive statistics, including mean, median, and standard deviation, were calculated for all predictive scores using Python libraries such as NumPy and SciPy. Correlation analysis between tools was conducted to identify patterns of agreement and discrepancies. Scatter plots, histograms, and heatmaps were generated using Seaborn and Matplotlib to visualize data trends and relationships among scores. Statistical thresholds for prioritization included:

- **CADD**: High scores (≥20) were deemed indicative of significant functional impact.
- **REVEL**: Low scores (<0.2) highlighted variants with reduced pathogenicity predictions.

- **SIFT**: Low scores (≤0.05) indicated likely damaging variants.

Variant Prioritization

Variants were prioritized based on a multi-criteria approach that combined CADD, REVEL, and SIFT scores. This integrative framework ensured that variants with high functional impact predictions across multiple tools were flagged for further investigation. Prioritization criteria included:

1. High CADD scores (≥20) indicating deleteriousness.
2. Discordant predictions between REVEL and other tools, particularly variants with high CADD but low REVEL scores.
3. Variants flagged by at least three predictive tools as deleterious.

Software and Workflow Reproducibility

Reproducibility was ensured through a structured computational workflow using Jupyter Notebooks. All analyses were conducted in a controlled environment, and detailed documentation of the code and methodology was maintained. The workflow included data cleaning, statistical analysis, visualization, and prioritization steps, enabling other researchers to replicate the findings seamlessly.

Visualization and Interpretation

Key insights were visualized through:

1. **Histograms**: Used to display the distribution of predictive scores, revealing trends and potential outliers.
2. **Scatter Plots**: Highlighted discrepancies between tools, such as CADD vs. REVEL scores.
3. **Heatmaps**: Illustrated correlations between scoring systems, emphasizing their concordance and differences.

RESULTS:

Descriptive Statistics of TP53 Predictive Scores

The descriptive analysis of SIFT, PolyPhen, CADD, REVEL, MetaLR, and Mutation Assessor scores revealed significant trends. CADD scores exhibited a high mean (18.9) and standard deviation, with over 75% of variants exceeding a score of 20, indicating their predicted deleterious nature. SIFT scores clustered toward 0, supporting the damaging predictions for a majority of variants.

In contrast, REVEL scores demonstrated a skew toward lower values, with only a subset of variants showing high scores (>0.5). MetaLR and Mutation Assessor scores were consistently high, suggesting agreement between these tools regarding the pathogenicity of variants.

Table 1 summarizes the descriptive statistics for all predictive scores.

Table 1: TP53 Variant statistics

	SIFT	PolyPhen	CADD	REVEL	MetaLR	Mutation_Assessor
count	3830	2240	3814	3806	3806	3788
mean	0.137287	0.38236	18.89748	0.599574	0.932898	0.593234
std	0.222717	0.403795	9.445288	0.279141	0.127273	0.253909
min	0	0	0	0	0	0
25%	0	0.014	13	0.3895	0.926	0.403
50%	0.03	0.164	23	0.609	0.978	0.618
75%	0.18	0.85125	26	0.865	0.99	0.844
max	1	1	35	0.982	0.999	0.927

Variants with the highest CADD scores (>20) are likely pathogenic or functionally significant. These scores emphasize their potential to impact protein function, requiring further clinical validation.Top 10 Variants with Highest CADD Scores are given in (table 2).

Table 2:Top ten CADD score

	Variant_ID	CADD
2744	rs1131691028	35
885	rs1131691028	35
2742	rs1131691028	35
887	rs1131691028	35
2743	rs1131691028	35
886	rs1131691028	35
3319	rs2073451331	33
3252	rs1057519996	33
281	rs11575996	33
280	rs11575996	33

Observations made from fig 1 are as follows;

- **SIFT**: Majority of scores cluster near 0, indicating predictions of likely damaging variants.
- **CADD**: Many variants have high scores (>20), suggesting strong functional impact.
- **REVEL**: Distribution skews toward lower values, though some scores are high.
- **MetaLR**: Scores are consistently high, implying these variants are predicted to be deleterious.
- **Mutation Assessor**: Concentration of scores around 0.5 indicates moderate to low functional impact.

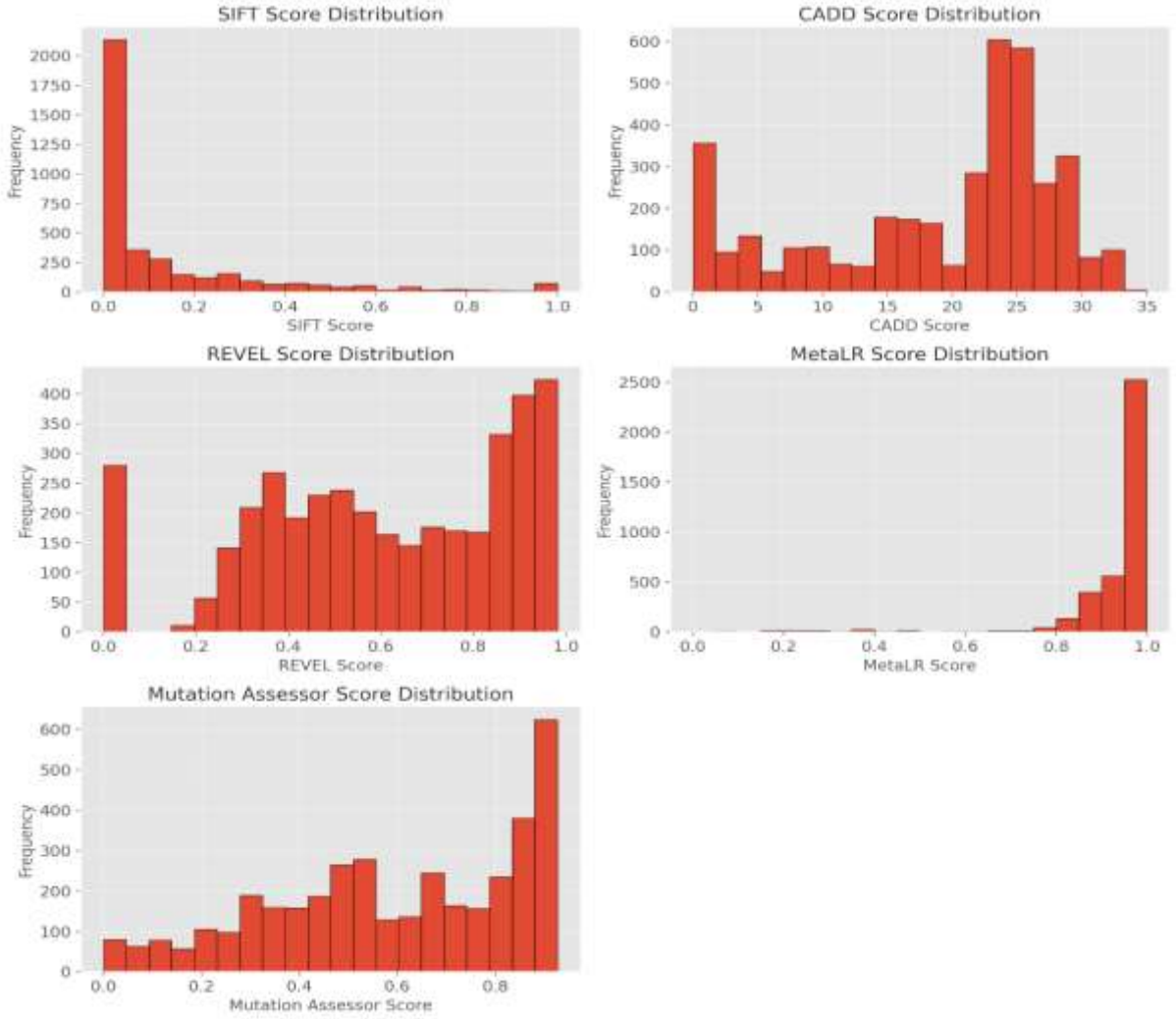

Fig 1: Score distribution chart

Correlation Analysis

A correlation matrix was generated to assess the relationships between predictive scores. Strong positive correlations were observed between CADD, MetaLR, and Mutation Assessor, reflecting consistent predictions of pathogenicity. However, REVEL showed weaker correlations with CADD (r = 0.64), suggesting discrepancies in their assessments of certain variants.**Table 3** presents the correlation coefficients between the scores.

There is a high positive correlation between **CADD**, **MetaLR**, and **Mutation Assessor** scores, suggesting that these tools consistently identify TP53 variants as likely pathogenic. For example, a correlation coefficient of **r = 0.73** between **CADD** and **Mutation Assessor** highlights strong agreement in predicting functional impact. This agreement increases confidence when multiple tools collectively predict a variant as deleterious, making these tools reliable for identifying clinically significant mutations in TP53-associated cancers.

The weaker correlation observed between **REVEL** and other tools (e.g., r = 0.64 with CADD) indicates inconsistencies in predictions. REVEL scores tend to prioritize more evolutionarily conserved mutations, which may miss certain structural or functional alterations captured by CADD. This discrepancy highlights the importance of integrating multiple predictive scores for a comprehensive assessment of variant pathogenicity.

Scatter plots of CADD versus REVEL further demonstrate the variability, with high CADD scores not always aligning with high REVEL scores(fig 2). Variants showing such discordance may require experimental validation.

High CADD and Low REVEL Variants: Variants with high CADD (>20) but low REVEL (<0.2) scores highlight inconsistencies in prediction tools. These variants may need functional validation to determine their significance.

Table 3: Correlations of scores

	SIFT	PolyPhen	CADD	REVEL	MetaLR	Mutation_Assessor
SIFT	1	-0.53782	-0.69335	-0.4843	-0.27534	-0.65281
PolyPhen	-0.53782	1	0.704026	0.737528	0.298285	0.717943
CADD	-0.69335	0.704026	1	0.649399	0.284605	0.735055
REVEL	-0.4843	0.737528	0.649399	1	0.346165	0.634111
MetaLR	-0.27534	0.298285	0.284605	0.346165	1	0.286357
Mutation_Assessor	-0.65281	0.717943	0.735055	0.634111	0.286357	1

Scatter Plot: A visualization of **CADD vs REVEL scores** shows that high CADD scores do not always align with high REVEL scores, reflecting differences in tool predictions(fig 2).

Priority Variants: I filtered variants with:

- **High CADD (>20)**: Strong predicted impact.
- **Low REVEL (<0.2)**: Low predicted pathogenicity.

- **Low SIFT (≤0.05)**: Damaging predictions.

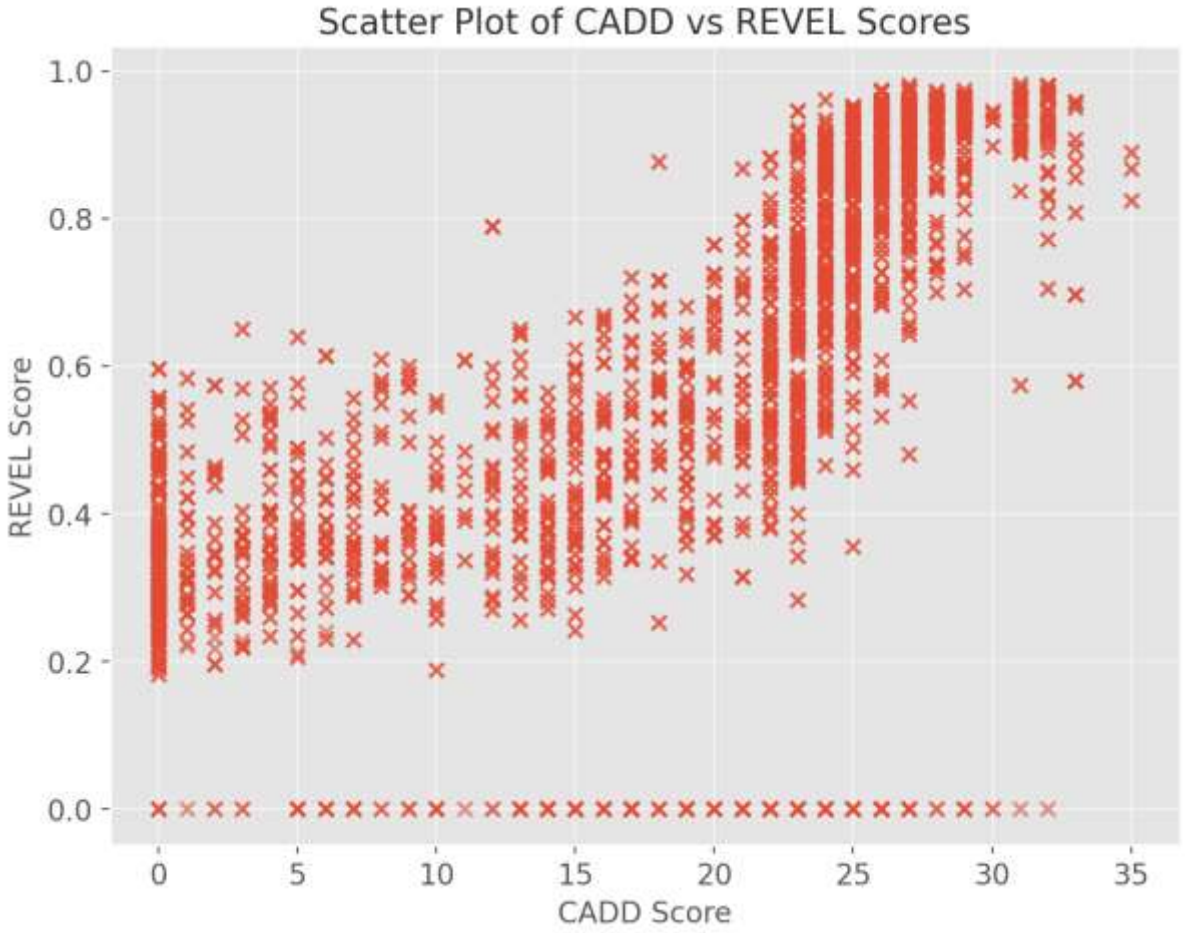

Fig 2: Scatter Plot of CADD versus REVEL Scores

DISCUSSION

This study provides a comprehensive analysis of TP53 variants using predictive scores to evaluate their pathogenicity. The results highlight the utility of tools like CADD and SIFT in identifying functionally significant variants. However, discrepancies between REVEL and CADD underscore the need for experimental validation to confirm predictions.[11,12]

TP53 mutations play a critical role in cancer progression by impairing DNA damage response mechanisms, leading to genomic instability. High-risk variants identified in this study may serve as biomarkers for cancer prognosis or therapeutic targets, particularly in cancers where TP53 alterations are prevalent.[13,14]

Combining many prediction systems increases confidence in spotting very harmful mutations. Still, limits such missing clinical annotations and inconsistent ratings draw attention to the difficulties in variant interpretation. Combining computational predictions with experimental data will help future research to improve the harmful categorization of TP53 mutations.

Emphasizing their importance in cancer biology, the study of TP53 mutations exposed important new directions on their functional influence. According to the CADD scores, several variations show high values (>20), usually connected with major negative impacts on protein function. Such high scores suggest that these mutations are probably functional, which helps to explain the characteristic of TP53 malfunction in cancer progression—that of reduced tumor suppressor activity.

In a similar vein, the SIFT scores concentrated mostly on 0, indicating a significant number of variations expected to be harmful. Further proving the toxicity of numerous TP53 variations, SIFT finds amino acid alterations most likely to influence protein structure or function. This

fits the way TP53 mutations disturb the p53 protein's capacity to control the cell cycle, DNA repair, and apoptosis, hence fostering unbridled cellular growth.

By contrast, the REVEL scores showed a slant toward lower values; less variations exceeded the high pathogenicity threshold (>0.5). This difference implies that REVEL, which combines many methods and evolutionary conservation, could minimize the functional effect of some TP53 mutations found by other prediction scores. Such arguments draw attention to the complexity of TP53 variant interpretation, in which computational methods might stress various elements of pathogenicity.

Strong links among CADD, MetaLR, and Mutation Assessor scores were found by further correlation study. These instruments repeatedly showed strong functional influence for certain versions, therefore supporting their dependability taken as a whole. Slower correlations between REVEL and CADD scores, however, revealed a collection of variations with contradicting expectations. High CADD scores but low REVEL scores define these discordant variations, which could be new or under-characterized mutations needing more functional confirmation.

The results highlight generally the need of combining many prediction techniques to precisely evaluate TP53 variations. High pathogenicity score variants across tools should be given top priority for additional research as they could be indicators for cancer diagnosis, prognosis, or therapy intervention. On the other hand, contradicting forecasts draw attention to shortcomings in present computational methods and the necessity of experimental validation to verify their clinical and biological relevance.

CONCLUSIONS

The need of include predictive scores to assess TP53 variations in cancer is shown by this work. This study helps the continuous efforts to categorize TP53 mutations for therapeutic uses by spotting high-risk variations and stressing differences across methods. These results need to be validated and their use in cancer diagnosis and precision medicine explored through more functional research.

This work underscores the importance of computational and statistical integration in variant analysis. By prioritizing high-risk TP53 variants using a multi-tool framework, this study advances our understanding of cancer-associated mutations and provides a foundation for experimental validation and clinical translation.

CONFLICTS OF INTEREST: NONE

FUNDING;NONE

REFERENCES

1. Hernández Borrero LJ, El-Deiry WS. Tumor suppressor p53: Biology, signaling pathways, and therapeutic targeting. *Biochim Biophys Acta Rev Cancer*. 2021;1876:188556. doi:10.1016/j.bbcan.2021.188556.
2. Kastenhuber ER, Lowe SW. Putting p53 in context. *Cell*. 2017;170:1062–1078. doi:10.1016/j.cell.2017.08.028.
3. Chen J. The cell-cycle arrest and apoptotic functions of p53 in tumor initiation and progression. *Cold Spring Harb Perspect Med*. 2016;6:a026104. doi:10.1101/cshperspect.a026104.
4. Olivier M, Hollstein M, Hainaut P. TP53 mutations in human cancers: origins, consequences, and clinical use. *Cold Spring Harb Perspect Biol*. 2010;2:a001008. doi:10.1101/cshperspect.a001008.
5. Silwal-Pandit L, Langerød A, Børresen-Dale AL. TP53 mutations in breast and ovarian cancer. *Cold Spring Harb Perspect Med*. 2017;7:a026252. doi:10.1101/cshperspect.a026252.
6. Giacomelli AO, Yang X, Lintner RE, et al. Mutational processes shape the landscape of TP53 mutations in human cancer. *Nat Genet*. 2018;50:1381–1387. doi:10.1038/s41588-018-0204-y.
7. Fischer NW, Ma YHV, Gariépy J. Emerging insights into ethnic-specific TP53 germline

variants. *J Natl Cancer Inst*. 2023;115:1145–1156. doi:10.1093/jnci/djad106.

8. Katsonis P, Wilhelm K, Williams A, Lichtarge O. Genome interpretation using in silico predictors of variant impact. *Hum Genet*. 2022;141:1549–1577. doi:10.1007/s00439-022-02457-6.
9. Larrea-Sebal A, Jebari-Benslaiman S, Galicia-Garcia U, et al. Predictive modeling and structure analysis of genetic variants in familial hypercholesterolemia: implications for diagnosis and protein interaction studies. *Curr Atheroscler Rep*. 2023;25:839–859. doi:10.1007/s11883-023-01154-7.
10. Ferla MP, Pagnamenta AT, Koukouflis L, et al. Venus: elucidating the impact of amino acid variants on protein function beyond structure destabilization. *J Mol Biol*. 2022;434:167567. doi:10.1016/j.jmb.2022.167567.
11. Ben-Cohen G, Doffe F, Devir M, et al. TP53_PROF: a machine learning model to predict impact of missense mutations in TP53. *Brief Bioinform*. 2022;23:bbab524. doi:10.1093/bib/bbab524.
12. Aguirre J, Padilla N, Özkan S, et al. Choosing variant interpretation tools for clinical applications: context matters. *Int J Mol Sci*. 2023;24:11872. doi:10.3390/ijms241411872.
13. Marei HE, Althani A, Afifi N, et al. p53 signaling in cancer progression and therapy. *Cancer Cell Int*. 2021;21:703. doi:10.1186/s12935-021-02396-8.
14. Marvalim C, Datta A, Lee SC. Role of p53 in breast cancer progression: an insight into p53 targeted therapy. *Theranostics*. 2023;13:1421–1442. doi:10.7150/thno.81847.