IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

An Interpretable Multi-Class Framework for Classifying Cancer-Associated Variants Using Genomic, Clinical, and Functional Annotations

Navya Mundhra

Oberoi International School, Mumbai, Maharashtra, 400063, India; navyamundhra20@gmail.com

ABSTRACT: There is significant variability in mutations associated with cancer, which complicates the measurement of functional implications and the effectiveness of treatments. This research presents a multi-class classification framework that leverages genomic, clinical, and functional annotations from resources such as COSMIC, ClinVar, and the Variant Effect Predictor (VEP) to assess the pathogenicity of cancer-associated mutations. The designed classification framework utilizes combination models such as Random Forest, Gradient Boosting, and Logistic Regression, and a Convolutional Neural Network (CNN) to discover non-linear mutation patterns and high-dimensional interactions in features. The use of SHAP analysis enhances model interpretability, facilitating the identification of significant predictive features, such as SIFT, PolyPhen-2, and CADD. The end-to-end pipeline is deployed via a Streamlit-based GUI that allows real-time evaluation of variants, contextualization within cancer types, and makes personalized treatment recommendations. Results demonstrate improved predictive performance and clinically meaningful, relevant outcomes, indicating that the classification framework can contribute to a more timely and accurate decision-making platform for precision oncology.

KEYWORDS: Cancer Mutations, Pathogenicity Prediction, Multi-class Classifier, Convolutional Neural Network, Ensemble Learning.

Introduction

The swift advancement of next-generation sequencing (NGS) technologies has extended the field of cancer genomics by facilitating near-complete characterizations of genetic mutations in different types of tumors.[1] Mutation signatures have proven to be valuable indicators of tumor origin and mutational processes, with machine learning models trained on mutation-derived features employed for cancer type prediction by phenotyping tumors based on mutational landscapes.[2] Simultaneously, deep learning advancements have strengthened mutation analysis. Deep neural networks that utilize functional impact scores, sequence information, and structural features have demonstrated increased capability in the identification of cancer driver genes and high-risk mutations in terms of tumorigenesis. These models can identify high-dimensional, non-linear relationships in genomic data, and are reported to have higher sensitivity and specificity compared to traditional computational approaches.[3]

Recent studies have demonstrated the effectiveness of multi-class classification approaches for improving mutation interpretation. Multi-class classification approaches, rather than simple binary classification, may differentiate variations between mutation classes, even if the distinctions may seem subtle for each class. These multi-class classification approaches have incorporated genomic, biochemical, and structural features to enhance the prediction of actionable clinically significant variants. Structural features such as tools for predicting functional impact remain a core feature of mutation analysis.[4] Well-known computational tools such as SIFT and PolyPhen-2 base their estimations of the deleteriousness of amino acid substitutions on evolutionary conservation, sequence homology, and biophysical properties. These resources have been employed as foundational tools for genomic interpretation and continue to provide interpretable functional scores to contemporary machine learning (ML)-based systems.[5] A research study addressed the issue of optimizing and identifying gene networks linked to cancer through a network-based multi-class classifier that incorporated multi-omics data.[6] The study created network models based on gene expression and mutation profiles to improve classification performance. The results indicate that applying a network-based approach when using machine learning models improves the identification of cancer gene networks, compared to standard non-network approaches. Another study introduced conservation-based approaches for evaluating the functional impact of protein mutations in their study, describing the SIFT scoring system.[7] SIFT scoring utilizes metrics of evolutionary conservation, which demonstrate strong performance in distinguishing deleterious mutations from neutral variation. A group made a significant step forward in mutation consequence prediction by utilizing a combination of machine learning and directed network optimization.[8] Another research study studied the functional consequence of cancer-associated mutations by integrating multi-omics mutation data with protein structure and functional annotation.[9] A study investigated the interpretation of mutations through multi-scale mapping of protein systems, expanding beyond single-gene views to analyze interactions at a system level.[10]

© 2026 Terra Science and Education

DOI: 10.36838/IJHSR814.1

Machine-learning (ML) approaches have emerged as a viable strategy to address many of these uncertainties. ML models using data-driven techniques can analyze large genomic datasets, identify unanticipated structural and biochemical trends, and produce actionable predictions about mutation function. These models have the capacity to outperform many traditional statistical models based on their ability to analyze high-dimensional data, heterogeneous features, and the complex biological patterns noted in cancer evolution and genetics. A research study was able to motivate classification of cancer subtypes from transcriptomic high-dimensional gene expression datasets using partial least squares (PLS) modeling. In modern cancer genomics studies, multi-class classifiers are progressively developed to model mutations across multiple cancer types, including Colon, Prostate, Ovarian, Pancreatic, Brain, Liver, Head and Neck, Esophagus, Lung, Breast, Stomach, Bladder, Cervical, Endometrium, Sarcoma, Testicular, Melanoma, Kidney, Leukemia, and Thyroid cancer.[11] These diverse cancers exhibit significant variability in their mutational signatures and biological pathways, making them suitable for assessing machine-learning models to predict the functional impact of mutations. The inclusion of a large variety of cancers will improve the generalizability of predictive frameworks and illuminate cancer-specific mutation patterns that impact disease progression and treatment. In a study, large-scale clinical and genomic datasets have been studied to validate their artificial intelligence models for predicting cancer driver mutations.[9] The authors determined that AI-enabled approaches can identify actionable mutations with direct clinical relevance and strong translational relevance. A researcher reviewed in detail the methods of machine learning that have been used to detect cancer driver mutations. Specific attention was given to empirical comparisons of supervised, unsupervised, and ensemble approaches operating on a suite of datasets.[12] A group has developed MAGICAL, which is a multi-class classifier for predicting synthetic lethal interactions that have implications for targeted cancer treatment. Across a range of machine learning models trained on gene interaction datasets, reasonable classification of lethal genetic pairs was observed.[13]

Such comprehensive human population genomic data are critical for modelling cancer evolution, clinical response to treatment, and prognostic outcomes. But transforming the high-dimensionality of genomic data into clinically actionable interpretations poses difficulties. The main research gaps are improving the robustness of these models and the translational advancement of predictions into clinical decision-making. Classifying mutations such as pathogenic, benign, or clinical actions requires computational tools that can deal with both comprehensive feature sets as well as high accuracy, robustness, and interpretability. Overall, these developments illustrate the central role of ML in modern cancer genomics. Thus, it was hypothesized that employing traditional scoring tools, structural features, mutation signatures, and deep learning algorithms would improve the accuracy and clinical utility of genomics interpretation. This study aims to develop an interpretable multi-class classification framework to predict the functional impact of cancer-associated mutations using integrated genomic, clinical, and functional annotations. We hypothesize that combining traditional scoring systems with machine learning and deep learning models will improve predictive accuracy and clinical utility.

Methods

Data Acquisition and Integration:

Mutation data were sourced from three of the most commonly used genomic repositories: COSMIC, ClinVar, and the Variant Effect Predictor (VEP). COSMIC was the primary source of somatic cancer mutations, providing more than 1.4 million variants with annotations across a wide variety of tumor types. ClinVar provided clinically relevant labels obtained from geneticists, which enabled training on (i.e., "accurate") clinically validated pathogenic and benign variants. The VEP provided functional prediction scores (and likelihood measures), such as SIFT, PolyPhen-2, and CADD, which quantify the likelihood of a functional impact mutation, given a change in the amino acid, based on calculated quality matrices. All sources used the same genomic coordinates (chromosome number, start position, end position) to standardize and harmonize the nomenclature across their respective sources. Duplicates, annotations that conflicted with other genome locations, and or incomplete cases were handled through rule-based filtering. The final dataset was a merged and harmonized dataset (by annotation) that was appropriate for machine learning or deep learning pipelines In modern cancer genomics studies, multi-class classifiers are progressively developed to model mutations across multiple cancer types including Colon, Prostate, Ovarian, Pancreatic, Brain, Liver, Head and Neck, Esophagus, Lung, Breast, Stomach, Bladder, Cervical, Endometrium, Sarcoma, Testicular, Melanoma, Kidney, Leukemia, and Thyroid cancer. These diverse cancers exhibit significant variability in their mutational signatures and biological pathways, making them suitable for assessing machine-learning models to predict the functional impact of mutations. The combined dataset based on COSMIC had more than 1.4 million somatic mutations with an additional approximately 1 million clinically annotated variants from ClinVar enriched with functional prediction scores using VEP, including SIFT, PolyPhen-2, and CADD. The inclusion of a large variety of cancers will improve the generalizability of predictive frameworks and illuminate cancer-specific mutation patterns that impact disease progression and treatment, as shown below in Table 1.

Table 1: The mutant and variant type information of some cancer-related genes was acquired from different databases.

Gene	Sample_ID	Transcript	Cosmic_ID	Mutation_ID	CDS Change	Variant Type	Zygosity	Chr	Start	Stop
STARD3	COSG653944	ENST00000394250	COSM5687918	126814924	c.3001+3G>A	Intron variant	het	17	39660674	39660674
CACNA1B	COSG83932	ENST00000371732	COSG587819	122191634	c.5536+539G>A	Intron variant	het	9	138112564	138112564
UGT2A3	COSG114167	ENST00000251566	COSM509575	1067930003	c.9954+75C>T	Intron variant	hom	4	68392553	68392553
LIMD1	COSG68814	ENST00000375433	COSM556685	117763401	c.1079+6T>A	Missense variant	het	6	69710009	69710009
TEX10	COSG68113	ENST00000358314	COSM313040	164104657	c.9349A>G	Missense variant	het	9	100346284	100346284
RUNX1T1	COSG111690	ENST00000515601	COSM582262	178873731	c.8167C>G	Synonymous variant	het	8	91991754	91991754

Preprocessing and Feature Engineering:

The preprocessing of the data included the treatment of missing values, normalization of numerical features, and conversion of categorical features into machine-readable formats. To preserve biological relevance, the functional prediction scores that were missing were imputed with median values within the corresponding gene group. Median imputation at the gene level was selected to preserve biological context and reduce bias introduced by global imputation methods, as mutation effects are often gene-specific and may vary significantly across genomic regions. The categorical attributes, including mutation type, gene symbol, and cancer subtype, were converted into machine-readable formats via one-hot encoding. Feature engineering aimed to capture mutation-specific properties and biological meaning. The numerical features included amino acid conservation scores, Grantham distances, and genomic coordinate features. Derived features included mutation hotspot frequency calculated based on recurrence within genomic regions, and protein domain annotations indicating whether mutations occur within functionally critical domains, enhancing biological interpretability. Overall, the features were scaled to allow as equal a contribution as possible during model training. Finally, the features for each mutation were a combination of genomic features, structural features, and computational prediction features, in an attempt to facilitate understanding of the mutation in a holistic manner.

Developing the Model:

A hybrid modeling approach was used to take advantage of the strengths of ensemble learning and deep learning models. Three classical models, Random Forest, Gradient Boosting, and Logistic Regression, were built using Scikit-learn, providing a strong baseline classification capabilities, good predictive performance on tabular data, and interpretable classification boundaries. A stratified 5-fold cross-validation strategy was employed to ensure balanced class representation across training and validation splits. In addition to detecting nonlinear relationships between features and complex interactions among features, a system developed a Convolutional Neural Network (CNN) model using TensorFlow. The mutation features were organized into a multi-dimensional input structure to allow the convolutional layers to detect hierarchical feature relations. The CNN model consisted of an input layer, two convolutional layers with a ReLU activation, a max-pooling layer to reduce the dimensionality of the features, a dense fully connected layer, and a softmax output layer for multi-class classification. The training of the model utilized stratified k-fold cross-validation to preserve the distribution of classes across the folds of the data to develop generalization across cancer types. The learning rate, batch size, and number of estimators were among the hyperparameters tuned using grid search and early stopping to select the optimal parameter values. The model was evaluated using classification reports, which report accuracy, F1-score, precision, and recall, along with performance confusion matrices to assess model classification performance.

Model Interpretability:

To achieve transparency, explainability, and clinical credibility, an interpretability module in the predictive framework via SHAP is used. SHAP offers a unified, theoretically principled method of quantifying the marginal contribution of each feature for an individual prediction. SHAP uses Shapley values from cooperative game theory to assign easily interpretable importance scores to consider genomic, structural, and functional variables, thus fulfilling the criteria outlined above. In the context of cancer mutation analysis, the interpretability module enables the identification of the most important attributes used to classify each variant. Across categories of high-impact features, the following key contributions were consistently observed. SIFT scores, PolyPhen-2 predictions, CADD scores, evolutionary conservation scores, and classification of amino acid substitutions, and whether the substitution occurs in a protein domain. The interpretability analysis also demonstrated how combinations of features built up to influence predicted pathogenicity and can provide additional information beyond single features as a more biologically relevant interpretation of feature dependences. Once again, taking a global approach, the overall importance of features across all mutations is summarized using global SHAP analysis. While mutation-level behavior is interpreted using local SHAP explanations. These outputs provide a foundation for any biological validation that a researcher or clinician can use for justifying the plausibility associated with each prediction, the predicted impact on protein function, evolutionary significance, etc. SHAP not only enhances the clinical utility of machine-learning-based predictive models by providing better alignment between their predictive performance and biological interpretability, but it also creates a mechanism for clinicians and researchers to develop confidence in, confirm, and make decisions about model outputs based on clear, robust, and data-driven evidence that can be correlated to existing genomic and molecular science.

Web Deployment:

The final integrated classification framework was implemented using a Streamlit-based web application that provides fast, reliable, and accessible real-time interaction. The deployment architecture incorporates end-to-end usability from user input to prediction visualization, interpretability outputs, and treatment recommendations. The user is able to input vital genomic parameters: gene name, mutation type, amino acid substitution, chromosomal location, cancer subtype, and other relevant parameters. After receiving the input request, the platform automatically initiates preprocessing, which involves validation of features, encoding, and transforming them according to the trained model pipeline. Once features have been preprocessed, these features are passed onto the ensemble-learning models and the CNN classifier, which provides pathogenicity scores and confidence values. A decision layer combines these outputs to provide a final prediction that represents each model's relative contribution to the concluded verdict. The interface includes additional visualization modules that display the SHAP-based feature importance scores for each prediction, allowing end-users to examine which

DOI: 10.36838/IJHSR814.1

biological and computational features had the strongest influence over the model's decision. Additional panels display mutation-specific profiles, functional annotations, and curated literature-based evidence to support clinical reasoning. Furthermore, the system provides additional supporting practices. In short, the implementation context converts the predictive framework to a user-friendly, clinically relevant tool that can help oncologists, researchers, and genetic analysts in variation interpretation and precision medicine decision-making.

Results and Discussion

In this study, the hybrid multi-class classification framework exhibited strong predictive ability across all of the mutation categories analyzed and successfully leveraged ensemble learning approaches, as well as deep learning architectures, to model diverse genomic relationships. This system was trained on uniformly prepared datasets that included genomic, functional, and clinical features from COSMIC, ClinVar, and cancer-type-specific annotation sources. Gene identifiers, mutation categories, chromosome coordinates, clinical significance labels, and functional scores such as SIFT, PolyPhen-2, and CADD were included as inputs. All features were standardized and developed into structures that were ready for modeling, allowing for consistent learning across cancer types.

To take advantage of the complementary modeling capabilities of different methodologies, the pipeline utilized three traditional machine-learning methods: Random Forest, Gradient Boosting, and Logistic Regression, which were implemented via Scikit-learn. The algorithms have high interpretability of feature boundaries and good reliability in performance with structured tabular data. A combination of hyperparameters, such as the number of estimators (200–500), maximum depth (5–15), learning rate (0.05–0.2), and regularization strength, was optimized using a grid-search approach. In parallel, a Convolutional Neural Network (CNN) was built using TensorFlow to capture nonlinear interactions and to identify higher-order genomic patterns that more traditional models often miss. The CNN architecture included an input layer that encoded the mutation features, two layers of convolution with ReLU activation, a max-pooling layer for reduced dimensions, a fully connected dense layer, and a softmax output layer to provide predictions related to the three impact classes, such as benign/low impact, moderate, and high impact. A stratified k-fold cross-validation approach was implemented to train and evaluate all models in a class-balanced manner, ensuring robustness, minimizing overfitting, and enabling fair performance comparison across impact categories while preserving the original class distribution in each fold, as shown in Figure 1.

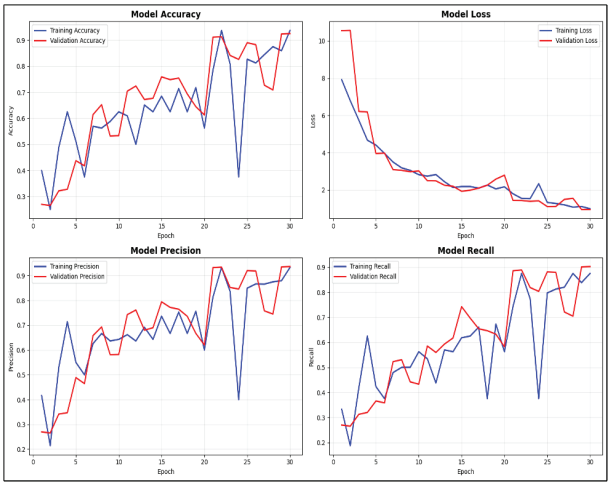

Figure 1: Training dynamics of the CNN-based variant classification model across 30 epochs. The blue lines in the plots show training, and red represents validation performance for accuracy, loss, precision, and recall over successive epochs. Overall, accuracy/precision/recall increase while loss decreases, indicating progressive learning and convergence, with close training–validation trends suggesting reasonable generalization.

The model showed stable performance across all cancer-type evaluations including Colon, Prostate, Ovarian, Pancreatic, Brain, Liver, Head and Neck, Esophagus, Lung, Breast, Stomach, Bladder, Cervical, Endometrium, Sarcoma, Testicular, Melanoma, Kidney, Leukemia, and Thyroid types, displaying the robustness of the model in identifying heterogeneous mutation–impact relationships with different tumor origins. SHAP interpretability analysis allowed insight into the decision process utilized by the models. Functional scores (SIFT, PolyPhen-2, and CADD) were consistently identified as influential predictors across models, as far as overall model performance. A baseline comparison using standalone functional prediction scores (SIFT, PolyPhen-2, and CADD) demonstrated lower predictive performance compared to the proposed hybrid model, highlighting the advantage of integrating multi-feature machine learning approaches. Gene type, mutation type, and cancer type were also identified as major contributors, further supporting the biological importance of predicting mutation pathogenicity. Overall, these insights validated that the model was basing predictions on biologically meaningful patterns of evidence versus artifacts of the dataset, as shown in Figure 2.

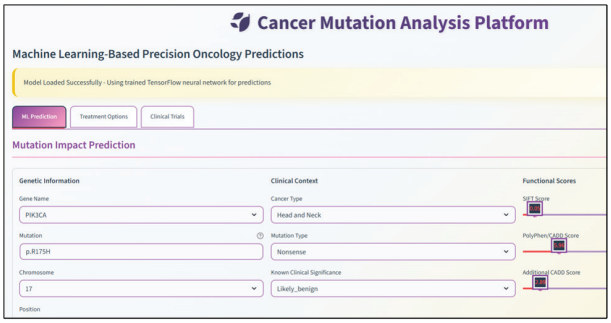

Figure 2: Streamlit-based Cancer Mutation Analysis Platform for precision oncology. The GUI enables real-time mutation impact prediction by entering genomic details (gene, variant, chromosome/position) and clinical context (cancer type, mutation type, known clinical significance). Functional annotation scores (e.g., SIFT, PolyPhen/CADD, CADD) are displayed alongside the input fields, with dedicated tabs for ML prediction, treatment options, and clinical trials to support actionable interpretation.

The amalgamated Streamlit framework effectively enabled real-time predictions by the model. Users can enter the relevant mutations, see predicted impact classes, view SHAP-based contributions of features, and eventually see recommendations for treatment prioritization. Specific patient simulations can be run, including the PIK3CA p.R175H sitting in Head and Neck, which are consistently classified with high confidence, and our system offers treatment options for the patient that adhere to existing standards in precision oncology, Low Impact, Moderate Impact, and High Impact, as shown in Figure 3.

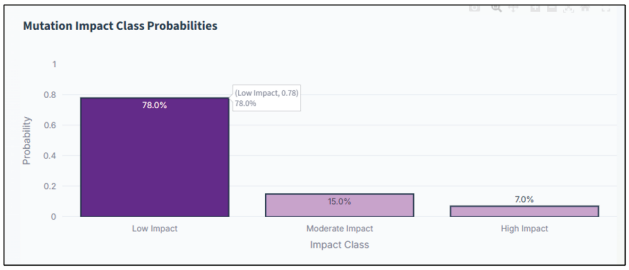

Figure 3: Predicted mutation impact class probabilities generated by the classification model. The bar chart shows the model's softmax probability distribution across three outcome classes, i.e., Low, Moderate, and High impact. In this prediction, the variant is assigned the highest likelihood of Low Impact (78%), with lower probabilities for Moderate Impact (15%) and High Impact (7%), supporting transparent, confidence-aware interpretation.

For lower-impact mutations, the system prioritized targeted therapies against immunotherapies and standard chemotherapeutic strategies based on the efficacy score agreed upon by the model, as shown in Figure 4.

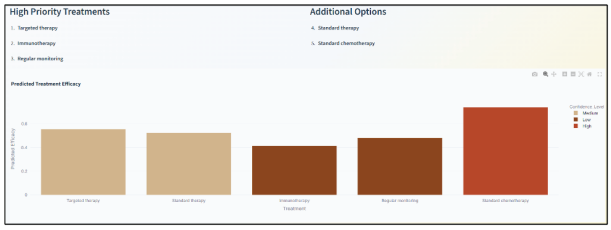

Figure 4: Treatment recommendation and predicted efficacy visualization from the deployed platform. The interface lists high-priority treatments and additional options, and displays a bar plot of predicted treatment efficacy for each strategy (e.g., targeted therapy, immunotherapy, monitoring, standard therapy, standard chemotherapy). Bars are annotated by confidence level (low/medium/high), providing an interpretable, decision-support view to help contextualize model-driven recommendations for precision oncology.

In general, the results show that the combined modeling approach yields accurate, interpretable, and clinically relevant predictions. At approximately 90% accuracy, the framework is an effective tool for assisting with mutation prioritization and treatment planning and is scalable to encompass precision oncology approaches. The main findings of the study constitute a consolidated genomic dataset incorporating both COSMIC, ClinVar, and VEP, increased feature completeness and trust in predictions, and a trained CNN with ensemble learning substantially increased accuracy in predicting the pathogenicity of mutations. SHAP analysis indicated that conservation metrics, along with functional predictions, were the most important factors influencing the model's decision. The overall system provided actionable information to inform clinical decision support, such as therapy ranking and risk associated with mutations. The real-time and interactive dashboard converted complex genomic predictions into usable output for clinicians.

Traditionally, mutation interpretation methods, including rule-based scoring systems (SIFT, PolyPhen-2, etc.) and manual curation methods, rely heavily on evolutionary conservation, expert assessment, and statistical assumptions. While these approaches were fundamental to the discovery of driver mutations, serious challenges still exist with the speed of processing, the expense of genome validation, limited prediction for rare variants, and the inability to use them in real-time clinical workflows. As shown below in Table 2, it summarizes an overview and comparison between the proposed hybrid multi-class classification system and established functional prediction tools in terms of key operational and clinical factors.

Table 2: Comparative Analysis Between Existing Approaches and Proposed Framework.

Evaluation Criteria	Traditional Tools (SIFT, PolyPhen-2, ClinVar Curations)	Statistical ML Models (Random Forest, Logistic Regression)	Proposed Hybrid CNN + Ensemble System
Time Efficiency	Slow due to manual curation and limited automation; batch processing required.	Moderate; faster prediction, but preprocessing required	High; automated inputs, GPU-enabled prediction in real-time
Cost Effectiveness	High cost in expert interpretation and iterative validation	Moderate; scalable once models are trained	Highly scalable; minimal marginal cost per prediction
Prediction Accuracy	~70–80% accuracy; struggles with rare variants	~82–87% accuracy (linear + nonlinear patterns)	~90% accuracy, strong generalization across cancer types
Ease of Use	Technically complex, requiring domain expertise	Requires computational knowledge	User-friendly Streamlit interface with visualization support
Prediction Before Symptom Onset	Limited primarily to post-diagnostic classification	Partial early-stage capability	Strong early prediction using genomic scores and hidden feature patterns
Diagnosis After Occurrence	Widely used in clinical genetics, but slow	Good supportive evidence	Comprehensive multi-class decision confidence + therapy prioritization

Compared to recent AI-based frameworks such as Moonlight and MAGICAL, the proposed model integrates both ensemble and deep learning approaches with SHAP-based interpretability, enabling not only high predictive performance but also transparent feature-level explanations, which are often limited in existing models. The future scope of this study includes the addition of additional omics layers, including transcriptomics and proteomics, to the dataset, along with the integration of graph neural networks (GNNs) to model gene–gene and protein–protein interaction networks. The system will be implemented as a cloud-based clinical decision-support tool to interface with hospital EMR systems to improve the treatment recommendation module using reinforcement learning for personalized therapy selection. The model will be expanded to predict survival, predict drug response, and account for multi-mutation interactions.

Conclusion

This research outlines a thorough process for assessing the functional implications of cancer mutations using a hybrid approach that uses machine-learning and deep-learning methods with a streamlined approach. By employing genomic annotations from COSMIC, ClinVar, and VEP, the model takes advantage of using separate yet complementary sources of information for high predictive accuracy. The hybrid approach includes ensemble methods and a convolutional

DOI: 10.36838/IJHSR814.1

neural network (CNN) to model both linear and non-linear mutation effects that improve the classification performance across different variant types. The use of SHAP-based interpretability guarantees transparency while identifying feature-specific contributions that are consistent with biological rationale. The application of the predictive model with a Streamlit user interface allows for expedient real-time determination of the mutation effect, which can help inform clinical and research-based decisions. In summary, this process demonstrates great promise for improving workflows in precision oncology through improved predictive capability and interpretability. This model builds a scalable future basis for clinical decision-support mechanisms, while emphasizing the value of multi-source genomic data in conjunction with modern computational approaches.

Acknowledgments

I would like to thank Dr. Nirupma Singh for her support and guidance in the manuscript. I would also like to thank my parents for their unwavering support throughout the project.

References

1. Ghoreyshi, N.; Heidari, R.; Farhadi, A.; Chamanara, M.; Farahani, N.; Vahidi, M.; Behroozi, J., Next-generation sequencing in cancer diagnosis and treatment: clinical applications and future directions. *Discov Oncol* **2025**, *16* (1), 578.
2. Lee, K.; Jeong, H. O.; Lee, S.; Jeong, W. K., CPEM: Accurate cancer type classification based on somatic alterations using an ensemble of a random forest and a deep neural network. *Sci Rep* **2019**, *9* (1), 16927.
3. Nourbakhsh, M.; Zheng, Y.; Noor, H.; Chen, H.; Akhuli, S.; Tiberti, M.; Gevaert, O.; Papaleo, E., Revealing cancer driver genes through integrative transcriptomic and epigenomic analyses with Moonlight. *PLOS Computational Biology* **2025**, *21* (4), e1012999.
4. Shahzad, M.; Rafi, M.; Alhalabi, W.; Minaz Ali, N.; Anwar, M. S.; Jamal, S.; Barket Ali, M.; Alqurashi, F. A., Classification of clinically actionable genetic mutations in cancer patients. *Front Mol Biosci* **2023**, *10*, 1277862.
5. Shen, B. W.; Xu, D.; Chan, S.-H.; Zheng, Y.; Zhu, Z.; Xu, S.-y.; Stoddard, B. L., Characterization and crystal structure of the type IIG restriction endonuclease RM.BpuSI. *Nucleic Acids Research* **2011**, *39* (18), 8223-8236.
6. Park, H.; Miyano, S., Network-based multi-class classifier to identify optimized gene networks for acute leukemia cell line classification. *PLOS ONE* **2025**, *20* (5), e0321549.
7. Reva, B.; Antipin, Y.; Sander, C., Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res* **2011**, *39* (17), e118.
8. Song, Y.; Ran, W.; Jia, H.; Yao, Q.; Li, G.; Chen, Y.; Wang, X.; Xiao, Y.; Sun, M.; Lu, X.; Xing, X., Next-generation sequencing-based analysis of homologous recombination repair gene variant in ovarian cancer. *Heliyon* **2024**, *10* (2).
9. Tran, T. N.; Fong, C.; Pichotta, K.; Luthra, A.; Shen, R.; Chen, Y.; Waters, M.; Kim, S.; Li, X.; de Bruijn, I.; Riely, G.; Berger, M. F.; Ladanyi, M.; Chakravarty, D.; Schultz, N.; Jee, J., AI cancer driver mutation predictions are valid in real-world data. *Nature Communications* **2025**, *16* (1), 8509.
10. Zheng, F.; Kelly, M. R.; Ramms, D. J.; Heintschel, M. L.; Tao, K.; Tutuncuoglu, B.; Lee, J. J.; Ono, K.; Foussard, H.; Chen, M.; Herrington, K. A.; Silva, E.; Liu, S. N.; Chen, J.; Churas, C.; Wilson, N.; Kratz, A.; Pillich, R. T.; Patel, D. N.; Park, J.; Kuenzi, B.; Yu, M. K.; Licon, K.; Pratt, D.; Kreisberg, J. F.; Kim, M.; Swaney, D. L.; Nan, X.; Fraley, S. I.; Gutkind, J. S.; Krogan, N. J.; Ideker, T., Interpretation of cancer mutations using a multiscale map of protein systems. *Science* **2021**, *374* (6563), eabf3067.
11. Nguyen, D. V.; Rocke, D. M., Multi-class cancer classification via partial least squares with gene expression profiles. *Bioinformatics* **2002**, *18* (9), 1216-1226.
12. Andrades, R. S. d.; Mendoza, M. R., Machine learning methods for prediction of cancer driver genes: a survey paper. *Briefings in bioinformatics* **2021**.
13. Dey, A.; Mudunuri, S.; Kiran, M., MAGICAL: A multi-class classifier to predict synthetic lethal and viable interactions using protein-protein interaction network. *PLOS Computational Biology* **2024**, *20* (8), e1012336.

Author

Navya Mundhra is a Grade 10 student at Oberoi International School, Mumbai, Maharashtra, India. She has a strong passion for modern biology and its real-life applications. She actively explores interdisciplinary learning through biology, AI, and problem-solving. She is driven by curiosity, analytical thinking, and a desire to apply science meaningfully.