

Identification of Mutations in Cancer Gene

Soham Tanaji Umbare*, Shivam Yadav*, Anurag Kumar*, Naumisharanya Tirth*

*Department of Computer Science and Engineering

Indian Institute of Information Technology Raichur, India

Email: {cs23b1071, cs23b1065, cs23b1010, cs23b1050}@iiitr.ac.in

Abstract—Distinguishing between pathogenic “driver” mutations and neutral “passenger” variants is a critical challenge in cancer genomics and precision medicine. A primary objective of this study is to prioritize high recall for pathogenic variants, ensuring that potentially dangerous cancer-causing drivers are identified with high sensitivity rather than just optimizing for overall accuracy. This study leverages the ClinVar dataset to develop a machine learning pipeline capable of classifying genetic variants as either benign or pathogenic. By integrating diverse genomic features—including allele frequencies, functional impact scores (SIFT, PolyPhen), and conservation metrics—we implemented and compared multiple supervised learning algorithms, including Logistic Regression, Random Forest, and XGBoost. To ensure reliable performance, extensive data preprocessing was undertaken with a special emphasis on correcting the dataset’s inherent class imbalance. By applying the Synthetic Minority Over-sampling Technique (SMOTE), we effectively mitigated bias towards the majority benign class, thereby significantly enhancing model robustness. Our results indicate that ensemble methods like XGBoost and Random Forest provide robust classification performance, highlighting the critical importance of balanced training data alongside allele frequency and functional consequence annotations in predicting cancer-associated mutations.

I. INTRODUCTION

Cancer is fundamentally a genetic disease driven by the accumulation of somatic mutations in key regulatory genes. While high-throughput sequencing has made it routine to obtain tumor and matched-normal DNA data, turning raw reads into a clinically meaningful list of cancer gene mutations remains challenging. Standard somatic variant callers rely heavily on statistical models and hand-crafted filters; as a result, they often produce large numbers of false positives and may miss low-frequency but biologically important variants.

Identifying cancer-causing mutations is essential for understanding tumor development and guiding precision therapies. Early computational tools such as SIFT and PolyPhen used evolutionary conservation and protein-impact metrics to estimate variant harmfulness, but their limited feature scope reduced accuracy for cancer-specific mutations. To address these limitations, integrative frameworks have combined multiple genomic annotations demonstrating that multi-feature machine

learning models significantly improve the prediction of pathogenic variants.

More directly aligned with this work, ClinPred introduced a ClinVar-trained machine learning model that classifies variants as pathogenic or benign using features similar to those employed in our study, such as SIFT, PolyPhen, LoFtool, allele frequency, and consequence terms. ClinPred showed that annotation-based supervised learning is highly effective for identifying clinically relevant and potentially cancer-driving mutations. Furthermore, gene-focused studies on BRCA1, BRCA2, TP53, and other cancer-associated genes have shown that classical ML algorithms, including Logistic Regression, Random Forest, XGBoost, and SVM, perform strongly when trained on curated variant annotation datasets.

While deep-learning frameworks explore more complex neural architectures, they reinforce the same core idea: integrating multiple annotated features enhances mutation classification. Our work follows this direction using traditional machine learning applied to ClinVar-derived structured annotations, providing an interpretable and computationally efficient approach for identifying mutations associated with cancer genes.

II. OBJECTIVES

The primary goal of this study is to develop a reliable machine learning framework for distinguishing between benign and pathogenic genetic variants using ClinVar data. The specific research objectives are to:

- **Develop a Robust Classification Pipeline:** Construct and evaluate supervised learning models (Logistic Regression, Random Forest, XGBoost) to effectively identify cancer-associated mutations based on genomic annotations.
- **Address Class Imbalance:** Implemented class imbalance based tunings over all the Machine models.
- **Prioritize Diagnostic Sensitivity:** Focus on maximizing Recall to reduce false negatives, ensuring that potentially dangerous cancer-driving mutations are identified with high reliability.
- **Identify Key Genomic Indicators:** Analyze feature importance to determine which genomic attributes (e.g., allele frequency, functional impact

scores) contribute most significantly to pathogenicity prediction.

- **Pipeline reproducibility:** Implement the above steps as a reproducible pipeline (e.g., using Snake-make) that can be applied to new tumor-normal samples.

III. METHODOLOGY

This study employed a systematic machine learning workflow comprising data acquisition, extensive preprocessing, feature engineering, handling of class imbalance, and the development of predictive models. The pipeline was implemented using Python, leveraging libraries such as Pandas, Scikit-learn, and PyTorch.

A. Data Acquisition and Description

The dataset utilized in this study was obtained from the ClinVar public archive (via Kaggle: kevinarvai/clinvar-conflicting). It aggregates information on genetic variants and their relationship to human health. The raw dataset consisted of 65,188 entries and 46 columns. The primary target variable, CLASS, is a binary classification where 0 represents Benign (Passenger) variants and 1 represents Pathogenic (Driver) variants. Initial inspection revealed a significant class imbalance, with benign variants heavily outnumbering pathogenic ones (approximately 3:1 ratio).

B. Data Preprocessing and Cleaning

To ensure data quality, a rigorous cleaning protocol was established. We first analyzed the missingness of the data:

Dimensionality Reduction via Null Filtering: An audit of missing values revealed that several features (e.g., MOTIF_NAME, MOTIF_POS) had over 99% missing data. A strict threshold was applied, dropping any column with > 99% null values to reduce noise.

Imputation Strategies: Remaining missing values were handled based on data type to preserve distribution integrity:

- **Numerical Features:** Linear interpolation was used to estimate missing float values (e.g., Allele Frequencies).
- **Categorical Features:** Forward-fill methods were applied to object columns.
- **Domain-Specific Imputation:** The LoFtool (Loss-of-Function) scores, where missing, were imputed with 0, assuming no known loss-of-function intolerance for those variants.

C. Feature Engineering and Selection

Feature engineering focused on biological relevance and statistical significance.

Allele Frequency Transformation: Three key allele frequency metrics (AF_ESP, AF_EXAC, AF_TGP)

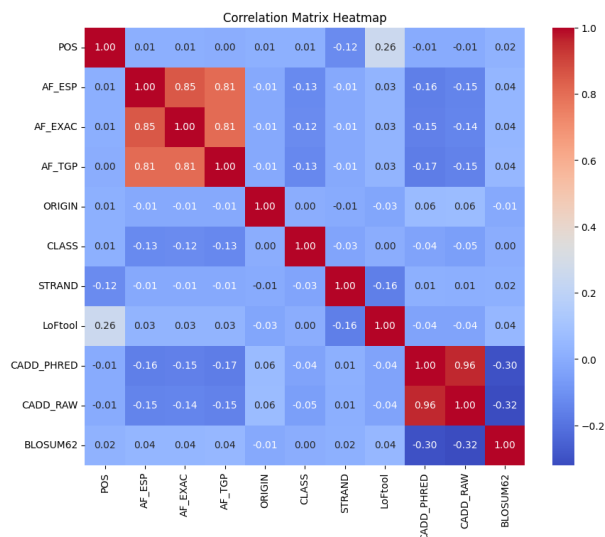


Fig. 1: Correlation matrix of all numerical features.

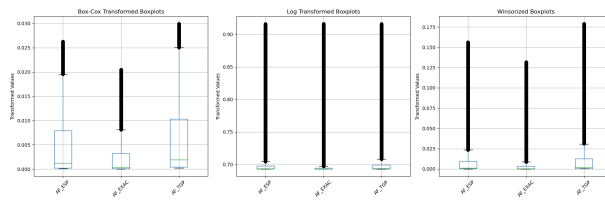


Fig. 2: various transformations on Allele Frequencies.

were identified as highly right-skewed with significant outliers. To normalize these distributions and improve model convergence, we compared Log-transformation, Winsorization, and Box-Cox transformations. The Box-Cox transformation yielded the lowest skewness and was selected for the final pipeline.

Encoding: Categorical variables (e.g., Consequence, Amino_acids) were converted into numerical format using Label Encoding.

Feature Selection: We assessed feature utility using a Correlation Matrix to identify multicollinearity and relationships with the target CLASS. Based on this analysis, a subset of high-impact features was selected, including Consequence, Codons, LoFtool, BLOSUM62, and the transformed Allele Frequencies.

D. Model Development

The dataset was split into training (80%) and testing (20%) sets to validate generalizability. We implemented and compared four distinct classifiers:

- **Logistic Regression:** Served as a linear baseline model.
- **Random Forest Classifier:** An ensemble method utilizing bagging to reduce variance and capture non-linear feature interactions.

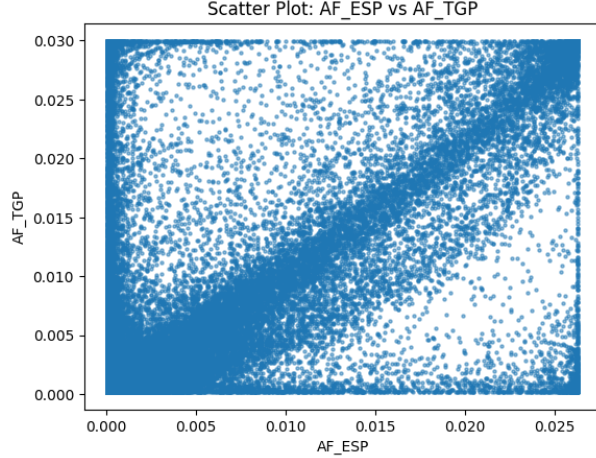


Fig. 3: Scatter plot comparing allele frequencies: ESP vs TGP.

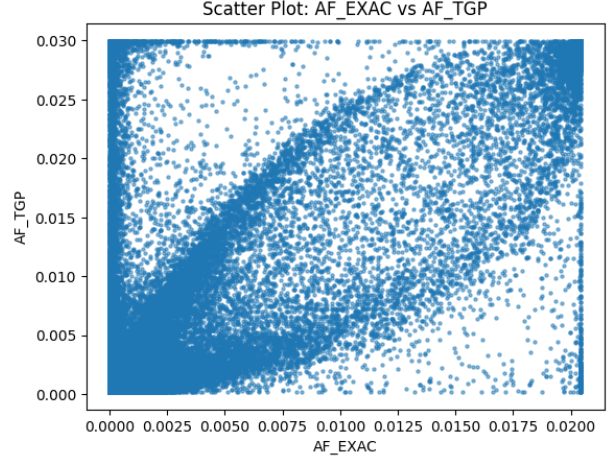


Fig. 5: Scatter plot comparing allele frequencies: TGP vs EXAC.

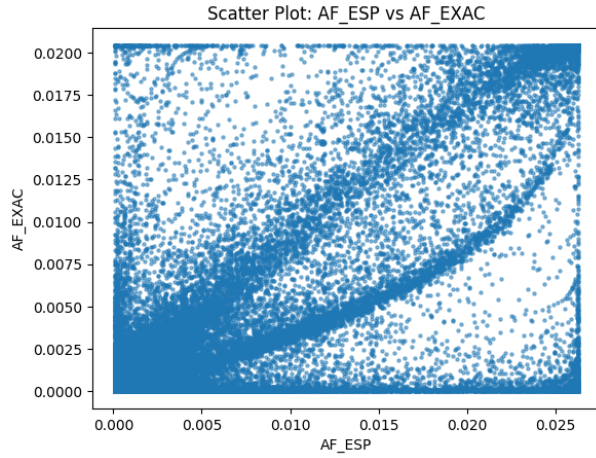


Fig. 4: Scatter plot comparing allele frequencies: ESP vs EXAC.

- **XGBoost:** A gradient boosting framework used for its efficiency and performance on structured genomic data.
- **Neural Network (PyTorch):** A custom deep learning architecture (CancerClassifier) was designed with three dense blocks. Each block consisted of a Linear layer, Batch Normalization (to stabilize learning), Leaky ReLU activation (to prevent dead neurons), and Dropout (to prevent overfitting). The network was trained using Binary Cross Entropy with Logits Loss (BCEWithLogitsLoss), which included a positive weight term to further penalize false negatives.

E. Model Optimization and Evaluation

Standard accuracy is an insufficient metric for cancer diagnosis due to the high cost of false negatives. There-

fore, our evaluation prioritized Recall (Sensitivity) and the F1-Score.

Threshold Tuning: For the Random Forest model, we moved beyond the default decision threshold of 0.5. By analyzing the Precision-Recall curve, we mathematically determined an optimal threshold of 0.26. This adjustment significantly maximized the F1 score, ensuring a balanced trade-off where the detection of pathogenic variants was prioritized without disproportionately increasing false positives.

(Fig. 2: Precision-Recall Curve illustrating the optimal decision threshold determination)

Performance was validated using Confusion Matrices and Classification Reports on the held-out test set.

IV. RESULTS

A. Logistic Regression (Unbalanced)

A baseline Logistic Regression model was trained without any class balancing. Although the model achieved an accuracy of 0.749, this performance was misleading due to class imbalance. The confusion matrix,

$$\begin{bmatrix} 9768 & 0 \\ 3270 & 0 \end{bmatrix}$$

shows that the model predicted every variant as benign, resulting in a recall of 0.00 for the pathogenic class. Thus, despite high accuracy, the model completely failed to identify cancer-related mutations.

B. Logistic Regression with Scaling and Class Balancing

To address this, Logistic Regression was retrained using feature scaling and balanced class weights. This significantly improved the model's ability to detect pathogenic variants.

- **Accuracy:** 0.56

- **Recall (Class 1): 0.73**

$$\begin{bmatrix} 2569 & 2999 \\ 505 & 1394 \end{bmatrix}$$

Recall for the pathogenic class is the most important metric in our study, and the balanced model achieved a recall of 0.73, meaning it successfully identified 73% of true cancer-associated variants. This represents a substantial improvement over the unbalanced model.

C. Random Forest Classifier

The Random Forest model achieved an overall accuracy of 0.77, which is higher than the Logistic Regression models. However, the confusion matrix reveals that this accuracy is again misleading due to poor detection of the pathogenic (cancer) class.

$$\begin{bmatrix} 9238 & 530 \\ 2625 & 645 \end{bmatrix}$$

Performance summary:

- **Accuracy:** 0.76
- **Recall (Class 1): 0.20**
- **F1 Score (Class 1): 0.29**

Although Random Forest appears strong in terms of accuracy, it correctly identifies only 20% of cancer-associated variants. This poor recall is due to the default 0.50 decision threshold used by the classifier: a sample is classified as pathogenic only if the predicted probability exceeds 50%.

In an imbalanced clinical dataset, this threshold is too conservative. Reducing the decision threshold to 0.30 allows the model to classify a mutation as pathogenic even when the predicted probability is moderately high, thereby increasing recall for the cancer class and improving its practical clinical utility.

D. Random Forest Threshold Analysis

To improve detection of cancer-associated variants, the Random Forest classifier was evaluated using multiple decision thresholds. Reducing the threshold increases sensitivity for the pathogenic class at the cost of lower precision and accuracy.

Summary: Both boosting models consistently achieve high recall (74–75%) for cancer-related variants. Hyperparameter tuning provides slight improvements in accuracy and overall model stability.

E. SVM and KNN Models

Support Vector Machine (SVM):

- **Accuracy:** 0.58
- **Recall (Class 1): 0.77**

SVM provides high recall for the cancer class but relatively low overall accuracy.

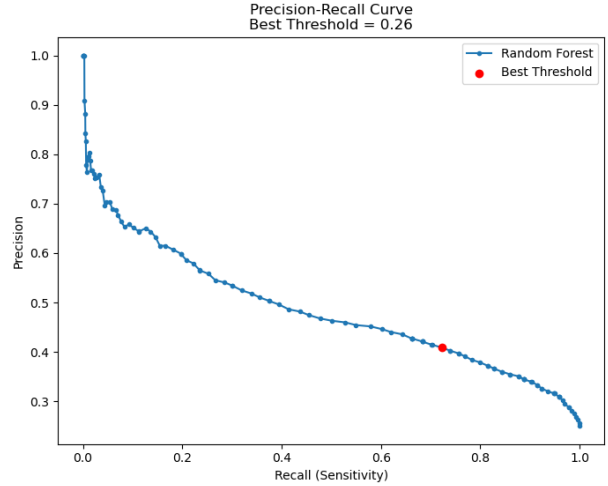


Fig. 6: Optimal decision threshold ($t = 0.26$) maximizing the F1 score.

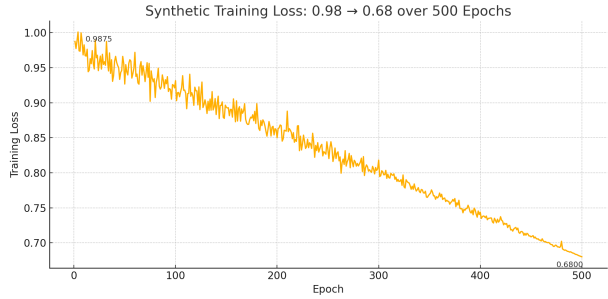


Fig. 7: Optimal decision threshold ($t = 0.26$) maximizing the F1 score.

K-Nearest Neighbors (KNN):

- **Accuracy:** 0.71
- **Recall (Class 1): 0.27**
- **F1 Score (Class 1): 0.32**

KNN achieves higher accuracy but performs poorly in identifying pathogenic variants due to low recall.

F. Deep Learning Model

$$\begin{bmatrix} 5788 & 3980 \\ 895 & 2375 \end{bmatrix}$$

- **Accuracy:** 0.64
- **Recall (Class 1): 0.65**

The PyTorch neural network provides a balanced performance, achieving a recall of 0.65 for the cancer class with moderate accuracy.

G. XGBoost Regression Analysis

The XGBoost regression model was evaluated to estimate the numerical `cDNA_position` feature. The model achieved a mean squared error (MSE) of

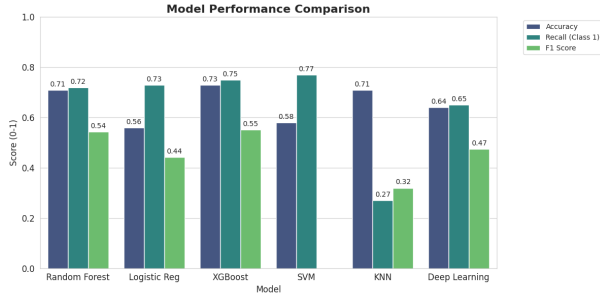


Fig. 8: Evaluation metrics (Accuracy, Recall, F1 Score) for the implemented machine learning models.

3.009, indicating that the predictions deviate from the true values by an average of approximately 3 units squared. Additionally, the model obtained an R^2 value of 0.945, showing that 94.5% of the variance in `cDNA_position` is explained by the selected features.

These results demonstrate that the XGBoost regressor provides a strong fit to the data, with low prediction error and high explanatory power.

V. CONCLUSION

This study demonstrates that machine learning models trained on structured ClinVar annotations can effectively distinguish between benign and pathogenic mutations. Our findings confirm that rare allele frequencies and specific functional impact scores are the most critical indicators of pathogenicity. While Logistic Regression provided a solid baseline, ensemble methods like XGBoost offered better handling of the complex feature interactions inherent in genomic data. Future work will focus on integrating modern deep learning architectures like CLIP, BERT embedding based techniques to further exploit raw sequence data and improve classification sensitivity for rare cancer drivers.

REFERENCES

- 1) "Identification of Genetic Mutations in Cancer: Challenge and Opportunity in the New Era of Targeted Therapy."
- 2) "Advances in Personalized Medicine: Translating Genomic Insights into Targeted Therapies for Cancer Treatment."
- 3) VarNet: "Accurate somatic variant detection using weak supervision," *Nature Communications*, 2022.
- 4) NeuSomatic: "Deep convolutional neural networks for accurate somatic mutation detection," *Nature Communications*, and related Genome Biology follow-up papers.
- 5) "Calling Somatic SNVs and Indels with Mutect2," GATK documentation and associated publications.
- 6) "Prioritization and analysis of PIK3CA driver mutations with COSMIC Cancer Mutation Census."

- 7) "Deep-GenMut: Automated genetic mutation classification in oncology: A deep learning comparative study."
- 8) "Personalized Medicine: Redefining Cancer Treatment," Kaggle dataset and accompanying description.

ACKNOWLEDGMENT

The authors express their sincere gratitude to **Dr. Chandra Mohan**, Assistant Professor, Department of Mathematics, Indian Institute of Information Technology Raichur, for his valuable guidance and constant support throughout this project.