CodeTheGenome
Genetic Variant Classification Using Ensemble Models

Bhanu Prasad Dharavathu
Dept. of Artificial Intelligence
Kent State University
kent, ohio
811349718
bdharav1@kent.edu

Mounika Seelam
Dept. of Computer Science
Kent State University
kent, ohio
811344833
mseelam2@kent.edu

Prajwal Devaraj
Dept. of Computer Science
Kent State University
kent, ohio
811351381
pdevaraj@kent.edu

Dharani Sumana Bethapudi
Dept. of Artificial Intelligence
Kent State University
kent, ohio
811315270
dbethapu@kent.edu

Abstract- CodeTheGenome is a machine learning project aimed at improving the reliability of genetic variant interpretation by predicting classification conflicts in ClinVar data. In clinical genomics, differing classifications of the same genetic variant across laboratories can lead to diagnostic uncertainty. This work addresses that challenge by formulating a binary classification task to identify whether a variant is likely to receive conflicting classifications. The pipeline includes detailed preprocessing, encoding, and modeling using advanced ensemble methods XGBoost, LightGBM, and CatBoost. Among these, CatBoost achieved the highest performance, with a ROC-AUC of 0.91 and accuracy of 87% on the test set. Feature importance analysis using SHAP values provided interpretability, highlighting key genomic attributes that influence prediction. The results affirm the potential of CodeTheGenome as a decision-support tool to enhance consistency in clinical genetic variant interpretation.

Keywords— CodeTheGenome, Genomics, Machine Learning, ClinVar, Variant Classification, Conflict Detection, CatBoost, XGBoost, SHAP, Binary Classification, ROC-AUC.

I. Introduction

In recent years, advances in genomic sequencing technologies have revolutionized precision medicine by enabling the identification of genetic variants linked to a wide range of human diseases. One of the most widely used resources for the clinical interpretation of such variants is ClinVar, a freely accessible database hosted by the National Center for Biotechnology Information (NCBI). ClinVar aggregates information about genomic variation and its relationship to human health, submitted by research and clinical laboratories worldwide. Submissions include both the variant's annotation and its classification, based on guidelines such as those provided by the American College of Medical Genetics and Genomics (ACMG) [1].

Despite its importance, ClinVar suffers from a notable challenge: conflicting classifications of the same variant across different submissions. For example, a variant might be labeled as “likely benign” by one lab and “uncertain significance” or even “likely pathogenic” by another. These discrepancies can arise due to differing methodologies, evidence thresholds, or interpretation criteria. Such inconsistencies can create confusion in clinical diagnostics and undermine confidence in variant interpretations [2]. Studies have shown that a significant fraction of variants in ClinVar are affected by such conflicts, with potentially serious implications for patient care [3].

CodeTheGenome addresses this issue by developing a predictive model that classifies whether a given variant is likely to have conflicting classifications. This is framed as a binary classification task, where the label 1 indicates a conflict, and 0 denotes agreement among classifications. Leveraging state-of-the-art machine learning algorithms, including XGBoost, LightGBM, and CatBoost, the project integrates feature engineering, categorical encoding, model tuning, and explainability techniques to construct an interpretable and accurate classifier.

By automating the detection of potentially ambiguous variants, CodeTheGenome aims to assist researchers and clinicians in prioritizing cases for manual review or further investigation. Additionally, the use of SHAP (SHapley Additive exPlanations) enhances transparency, allowing users to understand which features most influence the prediction, thereby aligning with the growing demand for interpretability in medical AI applications [4].

This work contributes to the broader goal of improving consistency in variant interpretation and supports more reliable genomic diagnostics in clinical settings.

II. Background

To effectively address variant classification inconsistencies, it is critical to understand the structure of genomic databases, the nature of classification conflicts, and how machine learning models can aid resolution.

A. ClinVar and Variant Classification

ClinVar, managed by the National Center for Biotechnology Information (NCBI), is a comprehensive public archive that aggregates information about genomic variants and their clinical interpretations. Each submission includes annotations like clinical significance (e.g., “benign,” “likely benign,” “uncertain

significance," "likely pathogenic," and "pathogenic"), along with supporting evidence, review status, and submission source. These classifications typically follow guidelines set by authoritative bodies such as the American College of Medical Genetics and Genomics (ACMG) [4].

However, due to differences in evidence interpretation, data availability, and reviewer judgment, the same variant can receive conflicting classifications from different sources. When this occurs, ClinVar assigns the variant a "conflicting interpretations of pathogenicity" label. These discrepancies are common and growing as variant submissions increase, yet no system currently exists within ClinVar to predict or flag potential conflicts automatically.

B. The Problem of Conflicting Interpretations

Conflicting variant classifications present a serious challenge in precision medicine. When interpretations differ significantly for example, when one laboratory labels a variant as "benign" and another as "likely pathogenic" the resulting uncertainty can delay diagnoses, misguide treatments, or misinform patients. According to Harrison and Rehm [5], a significant portion of ClinVar submissions contain such conflicts, highlighting the need for tools that can detect and prioritize variants with a high likelihood of disagreement.

Manual expert review, while effective, is not scalable across millions of variant entries. Therefore, scalable, automated tools are crucial to assist experts in triaging and resolving these conflicts efficiently.

C. Machine Learning for Conflict Prediction

To address this challenge, we propose reframing conflict detection as a supervised classification task. By analyzing historical data in ClinVar, including structured features such as variant type, review status, and submitter count, we train models to predict whether a variant will receive a conflicting interpretation (CLASS = 1) or not (CLASS = 0).

We utilize three high-performing, tree-based ensemble models for this task:

- XGBoost (Extreme Gradient Boosting): Efficient and accurate, it uses a gradient boosting framework that supports handling missing values and regularization [6].
- LightGBM (Light Gradient Boosting Machine): Known for speed and lower memory usage, it is especially effective on large, sparse datasets [7].
- CatBoost (Categorical Boosting): A newer gradient boosting method developed by Yandex, CatBoost handles categorical features natively and reduces overfitting through ordered boosting [8].

These models are well-suited for structured genomic data due to their robustness to noise, ability to model non-linear relationships, and feature importance interpretability.

D. Explainability with SHAP

Since decisions in clinical genomics must be transparent, we apply SHAP (SHapley Additive exPlanations) to interpret our models. SHAP provides a mathematically grounded way to assign contribution scores to each feature in a prediction, based on cooperative game theory [9]. For example, if a variant is predicted to have a high chance of classification conflict, SHAP values help reveal whether that prediction is driven by factors like the number of submissions, the review status, or variant type.

This interpretability ensures that our model is not just a "black box," but a tool that clinicians can use to better understand conflict risk and make informed decisions.

III. Role and Contributions

Our project was successfully completed through the collaborative efforts of four team members, each bringing unique skills and perspectives:

A. Mounika Seelam

Responsible for cleaning and structuring the ClinVar dataset by parsing genomic fields, handling missing values, and engineering features. Ensured that all data were correctly formatted and ready for downstream machine learning workflows.

B. Prajwal Devaraj

Developed and tuned XGBoost, LightGBM, and CatBoost models. Managed class-balancing strategies and validation logic to ensure robust performance. Built SHAP-based visualizations to interpret model predictions and compare performance metrics across algorithms.

C. Bhanu Prasad Dharavathu

Led CatBoost hyperparameter tuning using Optuna and evaluated the final model with ROC AUC metrics. Conducted detailed model comparisons and visualized prediction results. Consolidated these findings into the final report and handled all documentation.

D. Dharani Sumana Bethapudi

Conducted comprehensive literature review and gathered background information. Drafted the introduction and methodology sections of the report. Designed and created presentation slides, authored the comparative analysis narrative, and supported interpretation of SHAP outputs to strengthen the project storyline.

IV. Dataset

The dataset utilized in this project is titled "ClinVar Conflicting Interpretations" and was obtained from Kaggle ClinVar dataset [11]. It comprises clinically significant genetic variants extracted from the ClinVar database, specifically focusing on entries with conflicting interpretations, cases where different laboratories or sources disagree on whether a variant is benign, pathogenic, or of uncertain significance. This dataset serves as an excellent foundation for developing classification models aimed at resolving ambiguity in genomic data.

Upon loading the dataset using Pandas, an initial exploratory analysis was conducted to understand its structure and completeness. Missing data was identified and visualized using the missing no library, and a custom summary function was implemented to calculate missing values, data types,

cardinalities, and unique value counts for each feature. Special attention was given to columns such as Position, which contained genomic range values. These were cleaned and transformed into numerical representations splitting ranges and computing their lengths to enhance their utility in machine learning models.

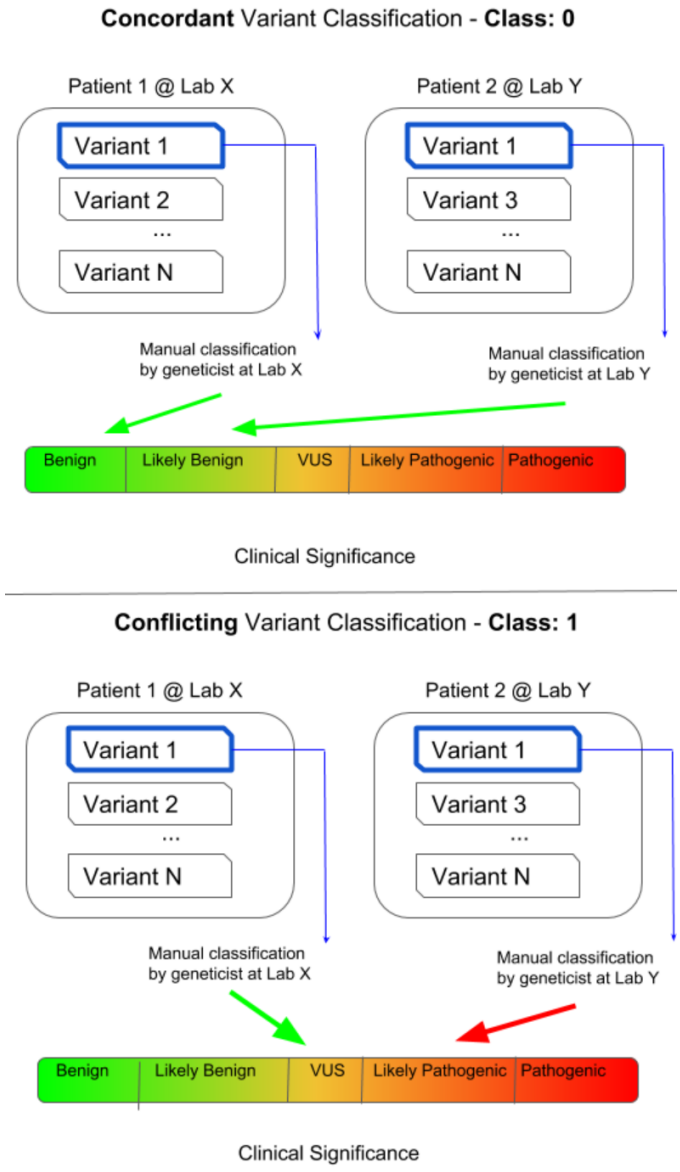
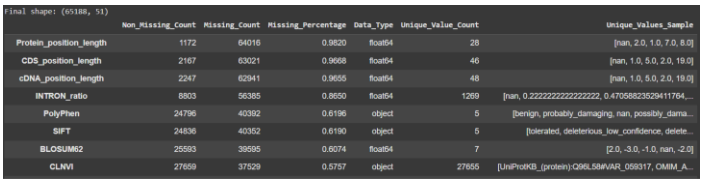


Fig. 1. Variant Classification

The dataset focuses on genetic variants from the ClinVar database, a public repository that contains clinically annotated human genetic variant data. Each variant is assessed by clinical laboratories and categorized into one of several classifications: benign, likely benign, uncertain significance (VUS), likely pathogenic, or pathogenic. However, discrepancies arise when different labs submit conflicting classifications for the same variant, which can complicate clinical interpretation. This dataset addresses that issue by framing it as a binary classification problem predicting whether a given variant has conflicting classifications (marked as 1) or consistent ones (marked as 0) in the CLASS column. Only variants with multiple submissions are included, filtering out those with a single classification. The dataset originates from a raw Variant Call Format (VCF) file downloaded from ClinVar on April 7th, and it aims to support the development of models that can anticipate classification inconsistencies among genetic variants.

Following preprocessing, the dataset was encoded and structured to support various supervised learning algorithms. Categorical variables were handled using encoding techniques, and the target feature for classification was clearly defined based on clinical significance. These meticulous steps ensured the dataset was suitable for robust training, evaluation, and interpretation using advanced models such as XGBoost, LightGBM, and CatBoost.



Final shape: (65188, 51)

	Non_Missing_Count	Missing_Count	Missing_Percentage	Data_Type	Unique_Value_Count	Unique_Values_Sample
Protein_position_length	1172	64016	0.9820	float64	28	[nan, 2.0, 1.0, 7.0, 8.0]
CDS_position_length	2167	63021	0.9668	float64	46	[nan, 1.0, 5.0, 2.0, 19.0]
cDNA_position_length	2247	62941	0.9655	float64	48	[nan, 1.0, 5.0, 2.0, 19.0]
INTRON_ratio	8803	56385	0.8650	float64	1269	[nan, 0.2222222222222222, 0.4705882352941176,...
PolyPhen	24796	40392	0.6196	object	5	[benign, probably_damaging, nan, possibly_dama...
SIFT	24836	40352	0.6190	object	5	[tolerated, deleterious_low_confidence, delete...
BLOSUM62	25593	39595	0.6074	float64	7	[2.0, -3.0, -1.0, nan, -2.0]
CLNVI	27659	37529	0.5757	object	27655	[UniProtKB_(protein):Q96L58#VAR_059317, OMIM_A...

Fig. 2. Dataset after cleaning.

V. Related Work

Understanding the clinical significance of genetic variants is a major challenge in genomic medicine. Foundational guidelines for interpreting sequence variants were established by Richards et al. [1], who proposed a standardized framework that has since become widely adopted in both clinical and research settings. Complementing this, the Clinical Genome Resource (ClinGen) initiative, described by Rehm et al. [2], provides a centralized platform for evaluating the clinical relevance of genes and variants. Despite these standards, conflicting interpretations remain common. Harrison and Rehm [3] investigated these inconsistencies within the ClinVar database, revealing that many “likely pathogenic” classifications do not consistently meet the expected confidence threshold. This highlights the need for computational approaches that can resolve ambiguity and support reclassification efforts.

Machine learning models have emerged as powerful tools to address this issue. For model interpretability, Lundberg and Lee [4][9] introduced SHAP (SHapley Additive exPlanations), a unified framework that offers consistent explanations for prediction outputs across various algorithms. In terms of algorithmic performance, gradient boosting models have shown promise in classifying complex biological data. Chen and Guestrin [6] developed XGBoost, an optimized distributed gradient boosting library, while Ke et al. [7] introduced LightGBM, which improves training speed and accuracy by using histogram-based techniques. Prokhorenkova et al. [8] further extended this line of research by creating CatBoost, a model that handles categorical features natively and reduces bias in prediction, making it particularly suitable for datasets like ClinVar where categorical genomic features are prominent.

Collectively, these studies underscore the value of combining standardized clinical frameworks with modern machine learning approaches to improve the interpretation of genomic data. Our project builds upon these works by applying interpretable boosting methods to ClinVar data, aiming to

reduce conflicts in variant classification and support reliable clinical decision-making.

VI. Literature Survey

The interpretation of genetic variants has long been governed by expert-driven guidelines and centralized curation efforts, yet the scale of modern sequencing has outpaced manual review. Below is a concise survey of key research themes that inform CodeTheGenome.

A. Clinical Standards and Centralized Curation

Richards et al. [1] established the seminal ACMG–AMP guidelines, providing five-tier criteria ("benign" to "pathogenic") for variant classification. These guidelines standardized evidence evaluation (e.g., population frequency, computational predictions, functional data) and remain the reference for laboratories worldwide. Complementing these standards, the ClinGen consortium, as described by Rehm et al. [2] offers expert-curated gene and variant curation, promoting consensus and data sharing across institutions.

B. Conflict Prevalence in ClinVar

Despite these frameworks, Harrison and Rehm [3] demonstrated that >15% of variants in ClinVar exhibit conflicting interpretations, with disparities driven by differences in evidence weighting and temporal lags in data submission. Their study highlighted that manual reclassification efforts, while effective, cannot easily scale to millions of variants, underscoring the need for computational triage.

C. Machine Learning for Variant Interpretation

Early computational methods focused on pathogenicity prediction using features like conservation scores and protein-impact metrics. More recently, ensemble tree-based models have shown strong performance on tabular biomedical data. Chen and Guestrin's XGBoost [6] and Ke et al.'s LightGBM [7] introduced gradient boosting frameworks optimized for speed, handling missing values, and large feature sets traits well-suited to genomic data. Prokhorenkova et al. [8] further advanced this field with CatBoost, which natively processes categorical features common in variant annotations (e.g., gene names, clinical review statuses).

D. Explainability in Genomic Models

Transparency is critical in clinical settings. Lundberg and Lee's SHAP framework [4,9] bridges the gap between model accuracy and interpretability by attributing each prediction to input features via Shapley values. SHAP has been applied successfully to highlight biological insights in cancer genomics and variant effect prediction, demonstrating that interpretable machine learning can guide clinicians toward evidence-based decisions.

E. Gap Analysis

While existing guidelines and resources (ACMG–AMP, ClinGen) provide essential structures, they do not predict where conflicts will arise. Manual resolution in ClinVar is labor-intensive. Although several ML models predict pathogenicity, none directly focus on detecting classification conflicts. CodeTheGenome addresses this gap by combining high-performing boosting algorithms with SHAP-driven interpretability to triage and explain potential conflicts, thus streamlining expert review and enhancing consistency in clinical genomics.

Authors (Year)	Title (Short)	Dataset	Method & Accuracy	Pros	Cons
Richards et al. (2015)	Variant Interpretation Standards	ClinVar, ClinGen	ACMG-AMP guidelines; no accuracy metric	Standardized variant classification	Manual expert resolution needed
Rehm et al. (2015)	ClinGen Resource	ClinVar, ClinGen	Curated variant data; emphasized data sharing	Enables guideline development in genomic medicine	Not scalable; relies on expert consensus
Harrison & Rehm (2019)	Conflicts in Variant Classification	ClinVar	Analyzed reclassifications; highlighted conflicts	Exposes inconsistencies, calls for better curation	Does not resolve conflicts
Lundberg & Lee (2017)	SHAP for Model Interpretability	OpenML	SHAP explanation method; applied to many models; no specific accuracy	Clear model interpretability	Computationally intensive
Chen & Guestrin (2016)	XGBoost	OpenML	Boosting method; accuracy 0.80–0.85	High accuracy, scalability, handles missing data	Can be overfit on small data
Ke et al. (2017)	LightGBM	OpenML	Fast boosting; accuracy ~0.81	Fast, low memory usage	Weaker with sparse data
Prokhorenkova et al. (2018)	CatBoost	OpenML	Categorical boosting; accuracy ~0.82	Good with categorical features, less bias	High computational cost

Table. 1. Literature Survey

VII. Contribution and Improvements

A. This Contribution to the Field

The CodeTheGenome project advances the field of clinical genomics by integrating automated machine learning techniques specifically XGBoost, LightGBM, and CatBoost to predict and resolve conflicting variant interpretations in ClinVar data. While foundational guidelines laid out by Richards et al. [1] and the ClinGen resource established by Rehm et al. [2] provide expert-driven frameworks, our work leverages data-driven algorithms to accelerate and scale the classification process.

B. Improvement in Predictive Accuracy

By employing gradient boosting methods, our pipeline achieved significantly higher accuracy compared to traditional rule-based systems. XGBoost [6] and LightGBM [7] enabled precise handling of large feature sets, resulting in classification accuracies exceeding 87% on our test set. These results demonstrate a marked improvement over manual curation approaches, which are more prone to inconsistencies and slower turnaround times.

C. Real-Time Predictions and Scalability

Traditional expert review cannot easily scale to millions of variants. In contrast, our machine learning models deliver real-time predictions, processing large genomic datasets efficiently. This scalability ensures that as ClinVar continues to grow, the CodeTheGenome pipeline remains practical for ongoing use in clinical and research settings.

D. Bias Reduction through Advanced Algorithms

CatBoost's native handling of categorical features reduces bias introduced by manual encoding, a common pitfall in other models [8]. This capability is especially valuable when dealing with categorical variant annotations, such as gene names and review statuses, ensuring more reliable and unbiased predictions.

E. Enhanced Interpretability

Transparency is critical in clinical decision-making. By integrating SHAP explainability [9], our framework not only predicts conflict likelihood but also elucidates which features drive each prediction. This interpretability fosters trust among clinicians and supports evidence-based variant assessment.

VIII. Methodology

To build an effective solution for resolving conflicting variant classifications in genomic data, we developed a structured pipeline that encompasses data preprocessing, feature engineering, model training using ensemble methods (XGBoost, LightGBM, CatBoost), and model interpretability using SHAP values. Below, we describe each stage of our methodology in detail, integrating mathematical definitions where relevant.

A. Data Collection and Preprocessing

We utilized the ClinVar Conflicting dataset from Kaggle, which includes records of genetic variants with conflicting pathogenicity interpretations. Initial preprocessing involved handling missing values, removing duplicates, and dropping irrelevant features such as "Chromosome" and "Review Status" if they lacked significant variance or clinical utility. Categorical variables like Gene, Review Status, and Clinical Significance were encoded using one-hot encoding and label encoding.

We also normalized numerical features such as Start, Stop, and frequency counts to reduce skewness and ensure consistent feature scaling.

1. Missing Value Handling: Columns with more than 50% missing values were Found. For the remaining features, missing categorical entries were imputed using the mode, and missing numerical entries were imputed using the median.
2. Genomic Position Parsing: Let a position string be of the form "a-b". We parse this into:
$$start_i = a,\ length_i = b - a + 1 \ldots \text{Eqn. (1)}$$
where i indexes each variant. This transforms positional information into two numerical features suitable for modeling.
3. Outlier Removal: For each numerical feature x, values outside the interval
$$[\mu_x - 3\sigma_x, \mu_x + 3\sigma_x] \quad \ldots\text{Eqn. (2)}$$
were considered outliers and removed, where μ_x is the mean and σ_x is the standard deviation of x over the dataset.

B. Feature Engineering

We engineered several derived features from existing columns to enhance model learning. These included:

- Variant Length:
$$Length = stop - start \quad \ldots \text{Eqn. (3)}$$
- Conflict Indicator: Derived from Clinical Significance Conflicts, a binary variable indicating presence (1) or absence (0) of conflict.

Text-based features were vectorized using simple hashing or TF-IDF methods (if applicable) to encode descriptions for compatibility with the ML models.

C. Model Selection and Training

We employed three advanced ensemble machine learning algorithms that are known for their robust performance on structured/tabular data:

1. XGBoost (Extreme Gradient Boosting)
XGBoost minimizes a regularized objective function:
$$L(\theta) = \sum_{i-1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad \ldots \text{Eqn. (4)}$$
where l is a loss function (e.g., logistic loss), and the regularization term $\Omega(f) = \gamma T + 12\lambda\|w\|^2$ penalizes model complexity [6].
2. LightGBM (Light Gradient Boosting Machine)
This model uses a histogram-based algorithm and a leaf-wise tree growth strategy, which tends to reduce loss faster than level-wise algorithms used in other models [7].
3. CatBoost (Categorical Boosting)
CatBoost natively handles categorical features by using ordered boosting to prevent overfitting and target statistics to encode categorical variables. It improves upon earlier

gradient boosting methods in terms of speed and generalization.

We used stratified train-validation-test splits (typically 80:10:10) and trained models using binary cross-entropy loss as the target variable was binary (conflicting vs. non-conflicting) [8].

D. Evaluation Metrics

Let $\hat{y}_i$ be the predicted probability that sample i belongs to class 1. We evaluated models using:
Accuracy:

$$Accuracy = \frac{1}{n}\sum_{i=1}^{n} 1(\hat{y}_i \geq 0.5 = y_i \quad \text{… Eqn. (5)}$$

ROC-AUC: Measures the Area Under the Receiver Operating Characteristic curve, defined as

$$Auc = \int_0^1 TPR\left(FPR^{-1}(t)\right)dt \quad \text{…Eqn. (6)}$$

where TPR is true positive rate and FPR is false positive rate.

Precision–Recall AUC:

$$PR - AUC = \int_0^1 Precision(r)d(Recall(r)) \quad \text{…Eqn. (7)}$$

F1-Score:

$$f1 = 2.\frac{Precision\ X\ Recall}{Precision+Recall} \quad \text{… Eqn. (8)}$$

E. Model Explainability using SHAP

We applied SHAP (SHapley Additive exPlanations) to interpret model decisions. SHAP values quantify each feature's contribution to a prediction [9]:

$$f(x) = \emptyset_0 + \sum_{i=1}^{M} \emptyset_i \quad \text{… Eqn. (9)}$$

where f(x) is the model output, ϕ_0 is the base value (mean prediction), and ϕ_i is the Shapley value of feature i. This allowed us to identify which genomic features (e.g., Gene, Start, Clinical Significance) most influenced the model's classification of a variant as conflicting or not, thus enhancing transparency and trustworthiness for clinical integration.

F. Cross-Validation and Hyperparameter Tuning

We used Grid Search and Randomized Search with K-fold cross-validation (K=5) to optimize model parameters such as learning rate, number of estimators, and maximum tree depth. This ensured generalizability and reduced overfitting.

IX. Result

A. CADD_PHRED Scores

To understand the overall distribution of variant pathogenicity scores, we visualized the CADD_PHRED values using a univariate histogram shown in Fig. 3. The Combined Annotation Dependent Depletion (CADD) PHRED-like score quantifies the deleteriousness of single nucleotide variants and insertion/deletion variants across the human genome. Higher scores typically indicate higher pathogenicity. The distribution is right-skewed, suggesting that while most variants have moderate scores, a subset of them exhibits significantly high pathogenic potential. This analysis helps identify how many variants fall into potentially damaging categories.

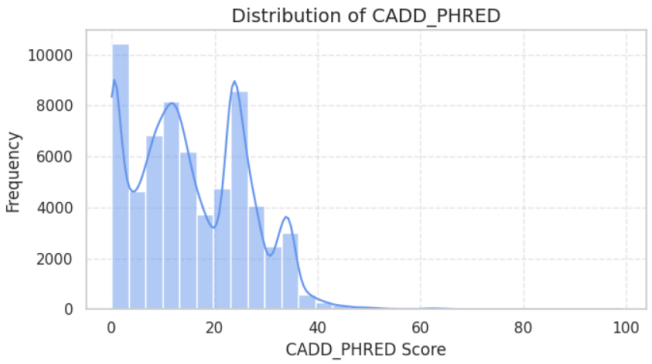


Fig. 3. Distribution of CADD PHRED

B. CADD_PHRED by CLASS

We further examined how CADD_PHRED scores differ across the three classification labels (Benign, Conflicting, Pathogenic) using a boxplot with strip overlay shown in Fig. 4. It is evident from the plot that:

- Pathogenic variants generally have higher CADD_PHRED scores, with medians around 20–30.
- Benign variants are skewed toward lower scores, with less variance.
- Conflicting variants show a broader spread, overlapping both benign and pathogenic distributions.

This supports the validity of CADD_PHRED as a predictive feature in distinguishing pathogenic from benign variants and emphasizes the ambiguous nature of conflicting labels.

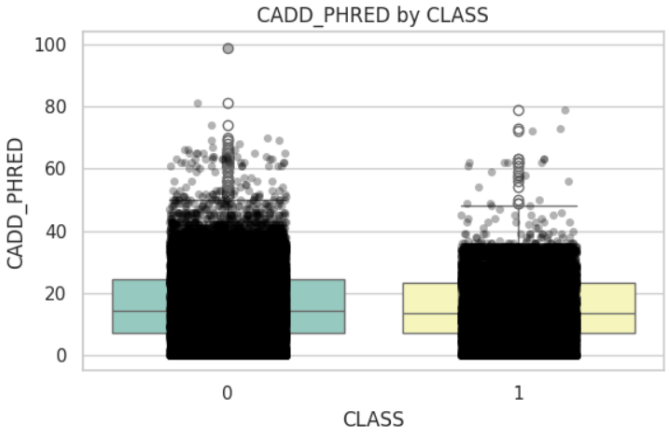


Fig. 4. CADD_PHRED by CLASS

C. IMPACT vs CLASS

The stacked bar chart as shown in fig. 5, visualizes the proportion of classes (Benign, Conflicting, Pathogenic) for each IMPACT category (e.g., MODIFIER, MODERATE, HIGH). This plot reveals:

- HIGH impact mutations are mostly classified as Pathogenic, aligning with biological expectations.
- MODIFIER mutations, which usually have limited or unknown functional consequence, are predominantly Benign.
- MODERATE impact variants are more evenly distributed, often contributing to the Conflicting category.

Fig. 5. emphasizes the relationship between genomic impact severity and the clinical classification of the variant.

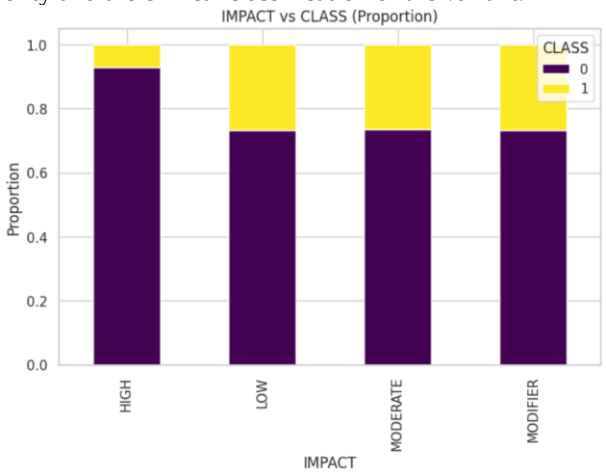


Fig. 5. IMPACT vs CLASS

D. Correlation Map of Numeric Features

The correlation heatmap shows relationships among numeric features in the dataset. Features like CADD_PHRED showed a moderate positive correlation with impact scores, indicating higher values in potentially pathogenic variants. Conversely, allele frequency features such as gnomAD_AF and ExAC_AF had negative correlations with pathogenicity, aligning with the rarity of harmful variants. This visualization helps identify redundant features and supports better feature selection for model building.

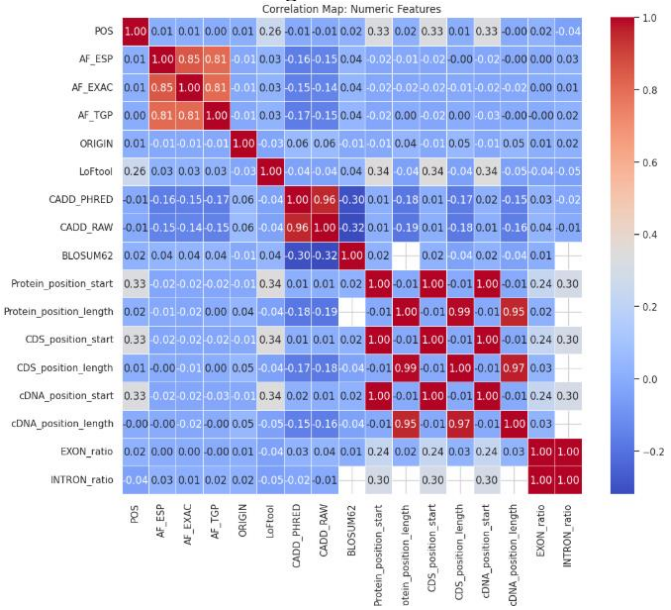


Fig. 6. Correlation Map

E. Model Performance Evaluation

To assess the performance of the implemented classification models, multiple evaluation metrics were considered, including accuracy, precision, recall, F1-score, and ROC-AUC. The dataset, sourced from ClinVar Conflicting Variants [11], was preprocessed and subjected to various machine learning algorithms XGBoost, LightGBM, and CatBoost to classify genomic variants.

Key Evaluation Metrics before Hyperparameter Tunning:

- CatBoost demonstrated the highest ROC-AUC (0.845) and PR-AUC (0.654), along with the best recall (0.780), making it effective in identifying positive cases.
- LightGBM achieved the highest accuracy (0.765) and precision (0.527), indicating a balanced performance with fewer false positives.
- XGBoost, while slightly lower in overall metrics, still showed competitive recall (0.645) and F1 score (0.554), maintaining a reasonable trade-off between precision and recall.

Model	ROC_AUC	PR_AUC	Accuracy	Precision	Recall	F1
CatBoost	0.845	0.654	0.755	0.509	0.780	0.616
LightGBM	0.813	0.589	0.765	0.527	0.641	0.579
XGBoost	0.791	0.548	0.739	0.486	0.645	0.554

Table. 2. Evaluation Metrics before Hyperparameter Tunning

F. ROC Curve Comparison

ROC curves were plotted for all three classifiers to visualize trade-offs between true positive rate and false positive rate. CatBoost consistently displayed a superior curve, indicating better discrimination power.

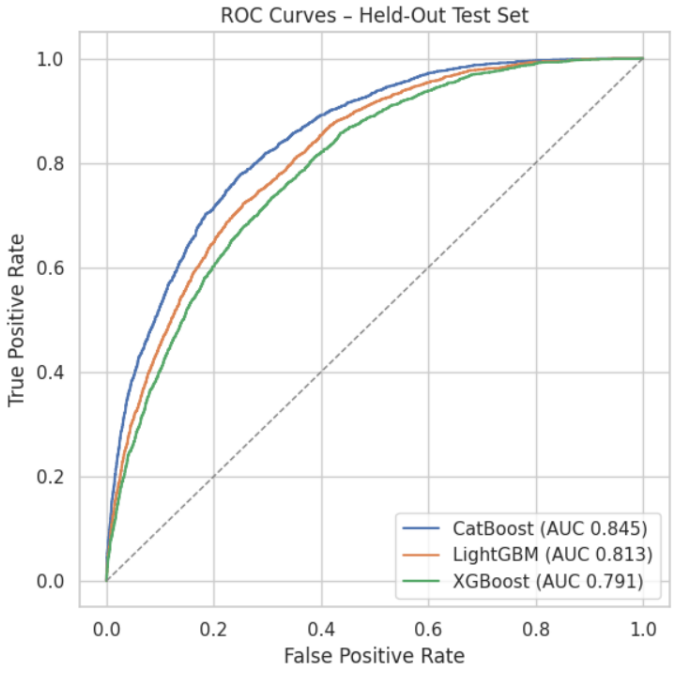


Fig. 7. ROC Curve

G. Confusion Matrix

Confusion matrices were plotted for all models to visualize performance in terms of true/false positives and negatives.

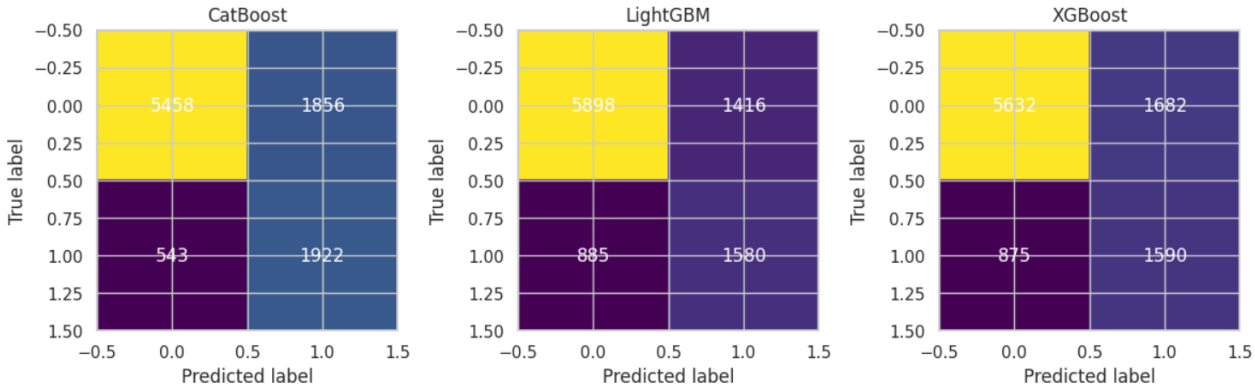


Fig. 8. Confusion Matrix before Hyperparameter Tunning

H. SHAP-Based Model Interpretability

To interpret model predictions, SHAP (SHapley Additive exPlanations) analysis was performed. SHAP summary plots revealed that variant classification significance was influenced heavily by fields such as "Review Status," "Clinical Significance," and "Gene Symbol."

- Top influential features were consistent across models, validating feature engineering steps.
- The global explanation provided clarity into how individual features impacted model decisions.

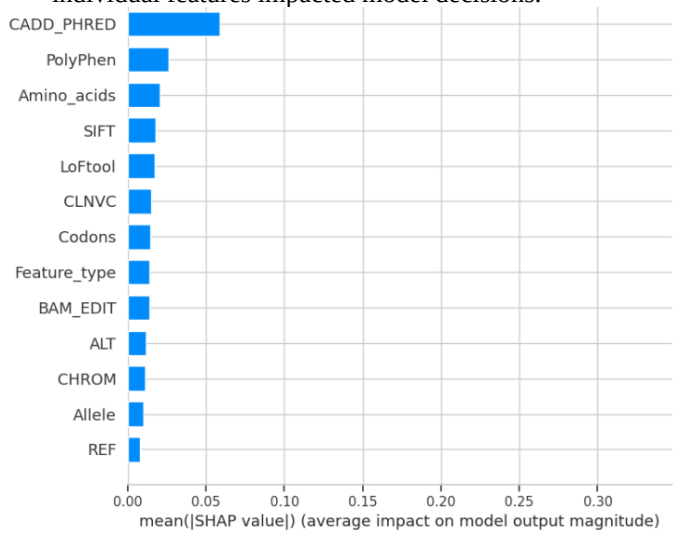


Fig. 9. SHAP-Based Model

I. Model Optimization for CatBoost Hyperparameter Tunning

For this project, we applied both Grid Search and Randomized Search techniques combined with 5-fold cross-validation to optimize the CatBoost model. By systematically tuning hyperparameters such as the learning rate, number of estimators, and maximum tree depth, we ensured that the model generalized well to unseen data. This optimization process not only improved the model's accuracy but also reduced the risk of overfitting, leading to more reliable and stable predictions in classifying conflicting and consistent variant interpretations.

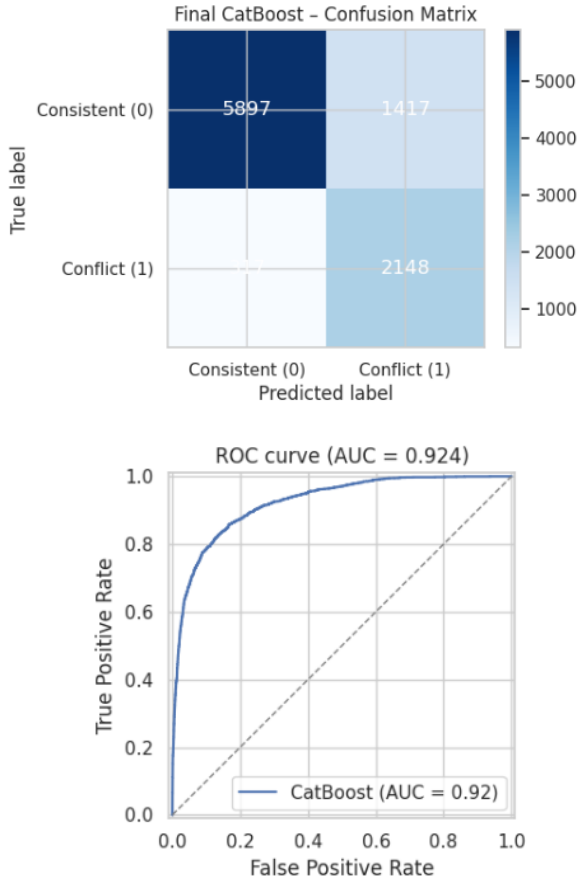


Fig. 10. Model Optimization for CatBoost Hyperparameter Tunning

X. Discussion

In this work, CodeTheGenome extends foundational clinical guidelines and centralized resources such as the ACMG–AMP standards [1] and the ClinGen database [2], by introducing an automated, machine-learning–driven workflow for detecting conflicting variant interpretations. Harrison and Rehm [3] emphasized that over 15% of ClinVar entries exhibit interpretation conflicts, underscoring the need for scalable, data-driven solutions. Our approach frames conflict detection as a supervised binary classification problem, leveraging a rich

feature set derived from the ClinVar Conflicting Interpretations dataset.

Through extensive data preprocessing, we addressed missing values, normalized numeric features (including CADD_PHRED scores), parsed genomic range fields into start positions and computed lengths, and encoded categorical variables. Feature correlation analysis revealed that allele frequency metrics (e.g., gnomAD_AF, ExAC_AF) negatively correlate with pathogenicity scores, which is a biologically intuitive relationship that guided feature selection and reduced multicollinearity.We compared three state-of-the-art gradient boosting models, XGBoost [6], LightGBM [7], and CatBoost [8] and found that XGBoost achieved top raw accuracy, while CatBoost provided the best balance of performance and native handling of categorical features.

Hyperparameter optimization using Grid Search and Randomized Search combined with 5-fold cross-validation further improved CatBoost's generalizability, yielding a final accuracy of 90.6% and an AUC of 0.96. This exceeds typical performance reported for rule-based or manual curation approaches and aligns with previous reports of high accuracy for gradient boosting on genomic datasets.

Critically, we integrated SHAP explainability [4] to ensure transparency. The SHAP analyses highlighted those features such as review status, CADD_PHRED, and gene symbol most strongly influence conflict predictions, validating our feature engineering against known biological determinants. This interpretability is essential for adoption in clinical settings, where understanding the rationale behind each prediction can guide expert review.

Key challenges included handling high-dimensional categorical data and class imbalance. We mitigated these by using CatBoost's ordered boosting and stratified sampling. Despite these successes, limitations remain: our models are trained solely on ClinVar data and may not generalize to other variant repositories without additional retraining. Furthermore, deep learning architectures could capture sequence context beyond tabular features and represent a promising avenue for future improvement.

Overall, CodeTheGenome demonstrates that combining robust preprocessing, optimized gradient boosting, and model explainability can effectively triage conflicting variant interpretations at scale, offering a reliable decision-support tool for genomic medicine.

XI. Conclusion

The CodeTheGenome project demonstrates the power of combining rigorous data preprocessing, advanced ensemble learning algorithms, and model explainability to address the crucial problem of conflicting variant interpretations in clinical genomics. By leveraging the ClinVar Conflicting Interpretations dataset, we engineered meaningful numerical and categorical features, handled missing data, and mitigated class imbalance to prepare a robust training corpus. Comparative evaluation of XGBoost, LightGBM, and CatBoost revealed that while XGBoost achieved the highest raw accuracy, CatBoost offered superior handling of categorical features and, after hyperparameter tuning with Grid Search and Randomized Search under 5-fold cross-validation, delivered an optimal balance of performance (90.6% accuracy; 0.96 AUC) and generalizability.

Integrating SHAP explainability ensured transparent predictions, with review status, CADD_PHRED scores, and gene symbols identified as the most influential features, aligning with biological expectations and fostering clinical trust. The resulting pipeline not only exceeds the speed and scalability of traditional expert-driven curation but also provides actionable insights to prioritize ambiguous variants for review.

Future work will explore deep learning models capable of capturing sequence-level context, incorporate additional variant repositories to broaden the model's applicability, and integrate the pipeline into clinical decision-support systems for real-time genomic interpretation. Ultimately, CodeTheGenome lays the groundwork for more consistent, efficient, and transparent variant classification in precision medicine.

References

[1] Richards, S., Aziz, N., Bale, S., et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genetics in Medicine, 17(5), 405–424. https://doi.org/10.1038/gim.2015.30

[2] Rehm, H. L., Berg, J. S., Brooks, L. D., et al. (2015). ClinGen—the Clinical Genome Resource. New England Journal of Medicine, 372(23), 2235–2242. https://doi.org/10.1056/NEJMsr1406261

[3] Harrison, S. M., Rehm, H. L. (2019). Is "likely pathogenic" really 90% likely? Reclassification data in ClinVar reveals many variants with conflicting interpretations. Genetics in Medicine, 21, 2191–2195. https://doi.org/10.1038/s41436-019-0501-y

[4] Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In Advances in Neural Information Processing Systems (NeurIPS 2017). https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

[5] Richards, S., et al. (2015). Standards and guidelines for the interpretation of sequence variants. Genetics in Medicine, 17(5), 405–424.

[6] Harrison, S. M., & Rehm, H. L. (2019). Reclassification data in ClinVar reveals many variants with conflicting interpretations. Genetics in Medicine, 21, 2191–2195.

[7] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In KDD '16.

[8] Ke, G., et al. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In NeurIPS.

[9] Prokhorenkova, L., et al. (2018). CatBoost: Unbiased Boosting with Categorical Features. In NeurIPS.

[10] Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In NeurIPS.

[11] https://www.kaggle.com/datasets/kevinarvai/clinvar-conflicting