Machine Learning to Detect Conflicting Clinical Classifications of Genetic Variants

Vallerie Cheng

Abstract

Identifying conflicting variants in the clinical classification of pathogenicity is important, as diagnoses directly affect the treatment plans for patients. Machine learning models can effectively categorize and analyze multidimensional data, and the incorporation of feature selection algorithms into these models allows us to identify relationships between clinical features that can contribute to conflicting clinical classifications of pathogenicity. We use the ClinVar dataset to address this need, which serves as a public archive for annotations of human genetic variants. In ClinVar, variants are manually categorized into different classes: benign, likely benign, uncertain significance, probably pathogenic, and pathogenic by researchers. However, there are inconsistencies in the annotations across clinical laboratories that can create confusion when assessing the impact of a variant on a patient's condition. This project proposes the development of a machine learning model trained on the ClinVar dataset to address this issue as well as feature selection to identify the properties of the variants that are most predictive of conflicting pathogenicity labels. The model leverages variant annotations such as genetic features, clinical data, and other critical information to identify patterns and relationships that harmonize conflicting classifications. We trained a random forest model and studied the importance of the input features using both tree-based and lasso feature selection. The five most significant features based on the tree-based feature selection, which inherently handles the nonlinear relationship between features, are (1) the score of the deleteriousness of variants, (2) allele frequencies emitted by ExAC, (3) Phred Scaled Score, (4) LoFtool's gene intolerance score, and (5) allele frequencies emitted by GO-ESP. The utilization of machine learning for identifying conflicting clinical classifications of genetic variants helps ensure precise and consistent interpretation of variants. This, in turn, plays a crucial role in improving clinical genomics, making diagnoses more accurate, and enabling personalized treatment options.

Introduction

The emergence of diseases and the evolution of virulent strains present significant challenges to the field of medicine. Accurate clinical classification of genetic variants is crucial for diagnosis, treatment, and various technical applications. However, the determination of pathogenicity in genetic variants often leads to divergent conclusions among scientists, influenced by numerous factors inherent to their respective distinct methodologies. Such inconsistencies in the classification of genetic variants result in patient assessments that can be severely influenced.

ClinVar is a freely accessible, public archive of reports on the relationships among human variations and phenotypes, with supporting evidence (14). It contains processed submissions reporting variants found in patient samples, assertions regarding their clinical significance, information about the submitter, and other investing data. We utilize ClinVar as a means to address our underlying scientific question because it facilitates the accessibility and communication of the connections established between variations in human characteristics and their impact on health, along with the historical context of the interpretations. Numerous features in the dataset affect the understanding of the pathogenicity of genetic variants. Moreover, researchers may reach disparate conclusions and subsequently submit conflicting reports concerning the classification of variants. This results in the inability to effectively devise treatment plans for patients while assessing the classifications of these variants. To address this problem, we seek to understand the most predictive features for predicting the pathogenicity of genetic variants to understand common variant features that lead to conflicting classifications. We trained a series of machine learning models which inherently perform feature selection, allowing us to prioritize the most significant features to identify a ranking of their weighted importance in detecting the presence of conflicting clinical classifications in genetic variants.

Related Work

Clinvar has been a widely used public archive to assess relationships between human variants and phenotypic characteristics. In recent years, numerous studies have contributed to developing machine learning models evolving around various aspects and the field of genomic variations (14).

Favalli et al. developed RENOVO, a machine learning-based tool that uses ClinVar data to classify genetic variants as pathogenic or benign by assigning pathogenicity likelihood scores (PLS) (4). The study showed that RENOVO effectively eliminates variants of uncertain significance (VUS) by providing a likelihood score for each variant, achieving an impressive F1 score of 0.95.

Larrea-Sebal et al. developed a machine-learning model called Mlb-LDLr to predict the pathogenicity of LDLr missense variants associated with familial hypercholesterolemia (7). By utilizing over 3,000 annotated variants from ClinVar, the model achieved an impressive AUROC value of 0.932 and predictive accuracy exceeding 90%.

Nicora et al. proposed using Penalized Logistic Regression and combining ACMG/AMP guidelines and variant annotation features to provide a probabilistic score for variant classification and prioritization (11). Their data-driven approach outperformed guidelines-based approaches and in silico prediction tools, successfully resolving more variants of uncertain significance (VUS), with the LR-A model achieving a mean accuracy score of 97.84%

Mahecha et al. compare three machine learning methods (Random Forest, Support Vector Machine, Multilayer Perceptron) for classifying VUS as Pathogenic or Non-pathogenic (9). The RF-based model outperforms the others, leading to the development of VusPrize, an open-source software tool for VUS prioritization.

Methods

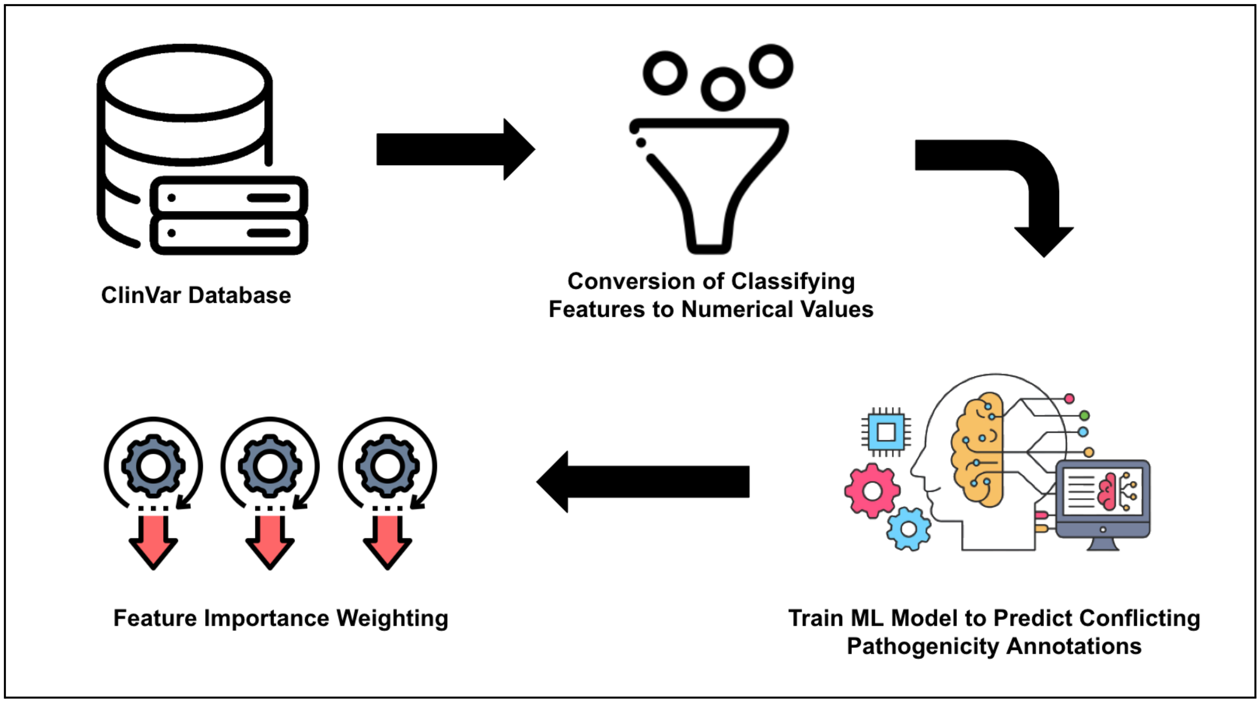
ClinVar Database
Conversion of Classifying Features to Numerical Values
Train ML Model to Predict Conflicting Pathogenicity Annotations
Feature Importance Weighting

Figure 1. Workflow for identifying important features related to predicting conflicting pathogenicity annotations of clinical variants.

Dataset Selection and Preprocessing

We used ClinVar, a publicly accessible archive of reports, that aggregates genomic information about clinical variants in the human genome (14). We preprocessed the dataset by transforming all input data into numerical values to enable effective computation and processing by the machine learning model.

Feature Representation and Imputation

We addressed missing data values (NaN) in the dataset by exploring various imputation methods. The first method we used was a simple imputer which substituted missing values with the mean of the respective column. We also used aKNN imputer, which averaged neighboring values uniformly. Finally, we used an iterative imputer which models features with missing values as functions of other features and utilizing these estimates for imputation.We evaluated several machine learning models with each imputation method. The Random Forest Classifier paired with iterative imputation yielded the highest performance, likely due to the ability of these methods to handle complex and nonlinear relationships between variables.

Data Balancing and Hyperparameter Optimization

Without data balancing, the model consistently predicted non-conflicting We therefore implemented data balancing by replicating rows in the training dataset to equalize the number of conflicting and non-conflicting labels. This adjustment prevented the model from favoring predictions of conflicting as the optimal choice.To fine-tune the model's behavior and enhance its predictive accuracy, we conducted hyperparameter optimization for the RandomForestClassifier. This process controlled the model's performance and minimized potential errors.

Feature Selection

We explored various feature selection methods within the machine learning model to identify the top-ranking features. This methodology is inspired by recent work in computational psychiatry, where the diagnostic feature space has been greatly reduced by applying feature selection to clinical questionnaires (7-8, 13, 14, 16-23). Initially, we removed features with low variance combined with tree-based feature selection, which resulted in a decrease in the F1 score. However, by adjusting the hidden layers of the neural network, we were able to observe an improvement in the score.

Additionally, we introduced Tree-based feature selection, utilizing the ExtraTrees Classifier, and Logistic Regression Feature Importances to derive feature coefficients essential for determining feature importance.

These thorough steps in dataset selection, preprocessing, model refinement, and feature selection were vital in constructing a robust machine learning model capable of effectively addressing conflicting clinical classifications in genetic variants within the ClinVar dataset.

Results

We display the top-ranking features using the extra cree classification method in Table 1 and the top-ranking features using logistic regression in Table 2. Eliminating the low variance method led to a noticeable F1 score improvement, rising from approximately 0.65 to 0.69. This observation indicated that removing features with low variance did not enhance the machine learning model, particularly in combination with other methods. Further improvement was achieved through L1-based feature selection, resulting in a modest increase in the F1 score, which ultimately reached 0.7014735870108884.

Feature	**Values**
Allele frequencies from GO-ESP (AF_ESP)	1.10685349e-01
Chromosome # (CHROM)	9.45870049e-02
Allele frequencies from ExAC (AF_EXAC)	1.49961013e-01
Allele frequencies from the 1000 Genomes Project (AF_TGP)	1.03560720e-01
Allele origin (ORIGIN)	9.26198754e-03
Type of consequence (Consequence)	2.91407351e-02
Impact modifier for the consequence type (IMPACT)	2.48498088e-02
Shortest distance from variant to transcript (DISTANCE)	5.83193638e-04
+ (forward) or - (reverse) (STRAND)	6.98192830e-03
Success or failure of edit using BAM file (BAM_EDIT)	1.56065897e-02
Loss of Function tolerance score (LoFtool)	1.16344545e-01
Phred-scaled CADD score (CADD_PHRED)	1.41649277e-01
Score of the deleteriousness of variants (CADD_RAW)	1.56045037e-01
Alignment scores for substituted amino acids depending on the particular acids substituted (BLOSUM62)	4.06787745e-02
Type of feature (Feature_type)	6.40371010e-05

Table 1. Coefficient values for ExtraTrees Classifier classification

Feature	Values
Allele frequencies from GO-ESP (AF_ESP)	-9.73811158e+00
Chromosome # (CHROM)	-3.68360365e-03
Allele frequencies from ExAC (AF_EXAC)	-4.36357376e+00
Allele frequencies from the 1000 Genomes Project (AF_TGP)	-1.18700834e+01
Allele origin (ORIGIN)	3.08764451e-03
Type of consequence (Consequence)	-3.29225979e-02
Impact modifier for the consequence type (IMPACT)	-2.06195892e-01
Shortest distance from variant to transcript (DISTANCE)	3.06206645e-04
+ (forward) or - (reverse) (STRAND)	-8.28703221e-02
Success or failure of edit using BAM file (BAM_EDIT)	-3.72074504e-03
Loss of Function tolerance score (LoFtool)	-4.27005917e-02
Phred-scaled CADD score (CADD_PHRED)	1.82172095e-02
Score of the deleteriousness of variants (CADD_RAW)	-1.14857934e-01
Alignment scores for substituted amino acids depending on the particular acids substituted (BLOSUM62)	-9.60016024e-03
Type of feature (Feature_type)	-2.71285911e+00

Table 2. Coefficient values for LogisticRegression Classification

Discussion

We delve into the feature importance analysis of our machine learning model, shedding light on the key determinants of conflicting or non-conflicting classifications of genetic variants. We have considered both ExtraTrees coefficients, which highlight feature importance in a broader context, and Logistic Regression coefficients, which signify whether a feature contributes to conflicting or non-conflicting predictions. Following is a ranking of the weighted importance of the selected features in the model.

CADD_RAW - Deleteriousness Score
The most crucial feature in our dataset for distinguishing conflicting from non-conflicting classifications is CADD_RAW. CADD_RAW represents the deleteriousness of a genetic variant, indicating how dangerous and pathogenic it is. A higher CADD_RAW value signifies greater pathogenicity. It is unsurprising that this feature is critical for classification. Logistic Regression further affirms this by assigning a negative coefficient to CADD_RAW, indicating that higher CADD_RAW values contribute to predicting non-conflicting classifications, aligning with the expectation that highly deleterious variants are less likely to be conflicting.

AF_EXAC - Allele Frequency
The second most important feature is AF_EXAC, which provides allele frequencies from the Exome Aggregation Consortium (ExAC) database (6). Allele frequency is crucial in understanding genetic diversity, and it aids clinical geneticists and biologists in assessing the impact of variants on diseases. Like CADD_RAW, AF_EXAC has a negative Logistic Regression coefficient, signifying its contribution to predicting non-conflicting classifications.

CADD_PHRED - Phred Scaled Score
CADD_PHRED, which incorporates a Phred scaled score, is pivotal in evaluating data quality and error rates in sequence analysis (2). A higher score indicates greater data quality and accuracy. In this case, a positive Logistic Regression coefficient suggests that higher CADD_PHRED scores contribute to predicting conflicting classifications. This feature also ranks high in ExtraTrees coefficients.

LoFtool - Gene Intolerance Score
LoFtool, a gene intolerance score based on loss-of-function variants (1) assists in assessing genic intolerance to functional variation. It ranks genes based on their intolerance to variation, helping identify variants susceptible to diseases. The Logistic Regression coefficient is negative, indicating its contribution to predicting non-conflicting classifications.

AF_ESP - Allele Frequencies from GO-ESP

AF_ESP provides allele frequencies from the Grand Opportunity Exome Sequencing Project (GO-ESP) database (10). While similar to AF_EXAC, it focuses on heart, lung, and blood disorders. Although less comprehensive than ExAC, AF_ESP still contributes to classification, as evident from its negative Logistic Regression coefficient.

AF_TGP - Allele Frequencies from 1000 Genomes Project
AF_TGP, which offers allele frequency data from the 1000 Genomes Project, is comparable to AF_EXAC and AF_ESP in providing insight into genetic variation. Its negative Logistic Regression coefficient indicates its role in predicting non-conflicting classifications.

CHROM - Chromosome Location
CHROM, denoting the chromosome location of a variant, assists in narrowing down potential disorders associated with chromosomal alterations. A negative Logistic Regression coefficient reinforces its contribution to predicting non-conflicting classifications. ExtraTrees ranks it high due to its ability to learn complex decision boundaries.

BLOSUM62 - Scoring Matrix
BLOSUM62, a scoring matrix used in protein sequence alignment (5), measures the probability of amino acid substitutions. Variants with higher probabilities of substitution are more likely to result in disease. As expected, it contributes to predicting non-conflicting classifications with a negative Logistic Regression coefficient.

Consequence - Variant Impact
Consequence provides information about the impact of variants on transcripts (3). It ranks lower in importance but still contributes to predictions, as indicated by its negative Logistic Regression coefficient.

IMPACT - Impact Modifier
IMPACT, an impact modifier feature, characterizes variants as HIGH, MODERATE, LOW, or MODIFIER based on their effects on proteins. While ranking lower in importance, it contributes to predicting non-conflicting classifications with a negative Logistic Regression coefficient.

BAM_EDIT - BAM File Editing
BAM_EDIT, indicating the success or failure of editing using BAM files (15), offers binary information that does not significantly impact classification. Nevertheless, it contributes to predicting non-conflicting classifications due to its negative Logistic Regression coefficient.

ORIGIN - Allele Origin
ORIGIN provides the allele's origin but holds limited influence on classification, as evident from its positive Logistic Regression coefficient, contributing to predicting conflicting classifications.

STRAND - Genetic Variant Strand
Similar to ORIGIN, STRAND provides information on genetic variant strands but ranks low in importance and contributes to predicting non-conflicting classifications due to its negative Logistic Regression coefficient.

DISTANCE - Variant to Transcript Distance
DISTANCE, offering the shortest distance from a variant to a transcript, has limited relevance to classification, as suggested by its positive Logistic Regression coefficient.

Feature_type - Feature Type
Lastly, Feature_type, detailing the genetic variant's feature type, ranks lowest in importance. Its contribution to classification is minimal, with a negative Logistic Regression coefficient.

In summary, the analysis of feature importance reveals that several key features significantly influence the classification of genetic variants as conflicting or non-conflicting. These features encompass various aspects of genetic data, including deleteriousness scores, allele frequencies, impact modifiers, and gene intolerance scores. The findings provide valuable insights for understanding the factors affecting clinical classifications in genetic variant datasets.

Limitations and Future Work

There were several limitations with our research effort that could have affected how accurate our results were. The main drawback was that we had to base our research only on one dataset. Due to this limitation, we were unable to analyze a wider range of features and may have missed important factors that could have affected our conclusions. In addition, it's possible that outside sources may have supplied pathogenicity labels that weren't present in the dataset, which could introduce biases into our research and result in areas where we don't fully understand. This limitation highlights how difficult it is to resolve missing data completely in our research.

To address the limitations identified in our research project and further enhance the accuracy and comprehensiveness of our findings, various future work improvements can be explored. First, by expanding the scope of data sources, we can increase the diversity of features that are available for researchers to analyze while reducing dataset-specific biases. Then, we can apply edge-cutting machine learning techniques that have shown to perform well in handling genomic data to future improve the accuracy of genetic variant classification. To verify the performance and results produced by the machine learning model, we can employ cross-validation techniques that conduct analyses to identify potential sources of biases. We aim to overcome the limitations found in this study and continue the advancement of genomic variant classification by addressing these future research applications, which will ultimately lead to more accurate and dependable assessments of variant pathogenicity for better clinical decision-making.

References

1. Academic.oup.com. (n.d.). https://academic.oup.com/bioinformatics/article/33/4/471/2525582
2. Combined annotation dependent depletion. CADD. (n.d.). https://cadd.gs.washington.edu/
3. Ensembl variation - calculated variant consequences. Calculated consequences. (n.d.-a). https://useast.ensembl.org/info/genome/variation/prediction/predicted_data.html#consequences
4. Favalli, V., Tini, G., Bonetti, E., Vozza, G., Guida, A., Gandini, S., ... & Mazzarella, L. (2021). Machine learning-based reclassification of germline variants of unknown significance: The RENOVO algorithm. The American Journal of Human Genetics, 108(4), 682-695.
5. Glossary. ROSALIND. (n.d.). https://rosalind.info/glossary/blosum62/
6. Karczewski, K. J., Weisburd, B., Thomas, B., Solomonson, M., Ruderfer, D. M., Kavanagh, D., Hamamsy, T., Lek, M., Samocha, K. E., Cummings, B. B., Birnbaum, D., The&nbsp;Exome&nbsp;Aggregation&nbsp;Consortium, Daly, M. J., &amp; MacArthur, D. G. (2017, January 4). The EXAC browser: Displaying reference data information from over 60 000 exomes. Nucleic acids research. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5210650/#B3
7. Larrea-Sebal, A., Benito-Vicente, A., Fernandez-Higuero, J. A., Jebari-Benslaiman, S., Galicia-Garcia, U., Uribe, K. B., ... & Martín, C. (2021). MLb-LDLr: a machine learning model for predicting the pathogenicity of LDLr missense variants. Basic to Translational Science, 6(11), 815-827.
8. Leblanc, E., Washington, P., Varma, M., Dunlap, K., Penev, Y., Kline, A., & Wall, D. P. (2020). Feature replacement methods enable reliable home video analysis for machine learning detection of autism. Scientific reports, 10(1), 21245.
9. Mahecha, D., Nuñez, H., Lattig, M. C., & Duitama, J. (2022). Machine learning models for accurate prioritization of variants of uncertain significance. Human Mutation, 43(4), 449-460.
10. NHLBI Grand Opportunity Exome Sequencing Project (ESP). (n.d.). https://esp.gs.washington.edu/drupal/
11. Nicora, G., Zucca, S., Limongelli, I., Bellazzi, R., & Magni, P. (2022). A machine learning approach based on ACMG/AMP guidelines for genomic variant classification and prioritization. Scientific reports, 12(1), 2517.

12. Tariq, Q., Daniels, J., Schwartz, J. N., Washington, P., Kalantarian, H., & Wall, D. P. (2018). Mobile detection of autism through machine learning on home video: A development and prospective validation study. PLoS medicine, 15(11), e1002705.
13. Tariq, Q., Fleming, S. L., Schwartz, J. N., Dunlap, K., Corbin, C., Washington, P., ... & Wall, D. P. (2019). Detecting developmental delay and autism through machine learning models using home videos of Bangladeshi children: Development and validation study. Journal of medical Internet research, 21(4), e13822.
14. U.S. National Library of Medicine. (n.d.). Clinvar. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/clinvar/
15. Variant effect predictor other information. Other information. (n.d.). https://useast.ensembl.org/info/docs/tools/vep/script/vep_other.html
16. Washington, P., Chrisman, B., Leblanc, E., Dunlap, K., Kline, A., Mutlu, C., ... & Wall, D. P. (2022). Crowd annotations can approximate clinical autism impressions from short home videos with privacy protections. Intelligence-based medicine, 6, 100056.
17. Washington, P., Kalantarian, H., Tariq, Q., Schwartz, J., Dunlap, K., Chrisman, B., ... & Wall, D. P. (2019). Validity of online screening for autism: crowdsourcing study comparing paid and unpaid diagnostic tasks. Journal of medical Internet research, 21(5), e13668.
18. Washington, P., Leblanc, E., Dunlap, K., Penev, Y., Kline, A., Paskov, K., ... & Wall, D. P. (2020). Precision telemedicine through crowdsourced machine learning: testing variability of crowd workers for video-based autism feature recognition. Journal of personalized medicine, 10(3), 86.
19. Washington, P., Leblanc, E., Dunlap, K., Penev, Y., Varma, M., Jung, J. Y., ... & Wall, D. P. (2020). Selection of trustworthy crowd workers for telemedical diagnosis of pediatric autism spectrum disorder. In Biocomputing 2021: proceedings of the Pacific symposium (pp. 14-25).
20. Washington, P., Park, N., Srivastava, P., Voss, C., Kline, A., Varma, M., ... & Wall, D. P. (2020). Data-driven diagnostics and the potential of mobile artificial intelligence for digital therapeutic phenotyping in computational psychiatry. Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 5(8), 759-769.
21. Washington, P., Paskov, K. M., Kalantarian, H., Stockham, N., Voss, C., Kline, A., ... & Wall, D. P. (2019). Feature selection and dimension reduction of social autism data. In Pacific Symposium on Biocomputing 2020 (pp. 707-718).
22. Washington, P., Tariq, Q., Leblanc, E., Chrisman, B., Dunlap, K., Kline, A., ... & Wall, D. P. (2020). Crowdsourced feature tagging for scalable and privacy-preserved autism diagnosis. medRxiv, 2020-12.
23. Washington, P., & Wall, D. P. (2023). A Review of and Roadmap for Data Science and Machine Learning for the Neuropsychiatric Phenotype of Autism. Annual Review of Biomedical Data Science, 6.