Prediction of Conflicts in Genetic Variant Classifications Through Machine Learning and Analysis of Predictive Features

NEEL GANDHI, Dartmouth College
SUNISHKA JAIN, Dartmouth College
DANIEL SHEN, Dartmouth College
XIAO YI WU, Dartmouth College
TEMILOLUWA PRIOLEAU, Dartmouth College

Abstract

With the revolution of next-generation sequencing, there has been a massive influx of genomic data. However, clinicians have struggled to keep up with the influx, resulting in the issue of conflicting classifications of the pathogenicity of genetic variants. Although the American College of Genetics and Genomics (AMCG) has created standardized guidelines for classification of variants, conflicting classifications are still quite common and can have huge impacts on a physician's plan for patient care. Previous works have not attempted to investigate which features are most relevant for the ultimate classification decision. Thus, we aim to not only create a model that performs well in all relevant metrics, but also investigates which features are most relevant for efficient classification performance. Using a filtered dataset from ClinVar, a publicly available archive on genetic variants, various methods such as one hot encoding, categorical variable assignment, and conversion between data types were used to generate feature sets. From there, we used the Synthetic Minority Oversampling Technique (SMOTE) to combat the issue of an imbalance dataset and feature normalization to combat the large differences in magnitude across features. After analyzing multiple machine learning models, the three best performing models were selected, two were tuned, and then the three were inputted into a unified stacked ensemble model that outperformed any individual model with efficient performance parameters. Following these results, a detailed analysis was carried out to understand the features which had the largest bearing on the binary classification through tools such as local interpretable model-agnostic explanations (LIME), SHapley Additive exPlanations (SHAP), permutation feature importance, and partial dependence plots (PDP). From these analyses, we found the features that play a large role in the classification of genetic variants. The results from our study contribute to the field of computational genetics with a novel stacked ensemble approach that achieved excellent accuracy, sensitivity, specificity, and Area Under the Receiver Operating Characteristics (AUROC). In addition to this, we extended the understanding of which features within ClinVar are most salient with regards to classification.

CCS Concepts: • **Applied computing** → **Computational genomics**; • **Computing methodologies** → *Reasoning about belief and knowledge*; *Ensemble methods*.

Additional Key Words and Phrases: ClinVar, Stacked ensemble machine learning model, genetic variant classification, Explainable AI

Authors' addresses: Neel Gandhi, Dartmouth College, neel.j.gandhi.gr@dartmouth.edu; Sunishka Jain, Dartmouth College, sunishka.jain.gr@dartmouth.edu; Daniel Shen, Dartmouth College, daniel.w.shen.23@dartmouth.edu; Xiao Yi Wu, Dartmouth College, xiao.yi.wu.gr@dartmouth.edu; Temiloluwa Prioleau, Dartmouth College, Temiloluwa.O.Prioleau@Dartmouth.edu.

ACM Reference Format:
Neel Gandhi, Sunishka Jain, Daniel Shen, Xiao Yi Wu, and Temiloluwa Prioleau. 2018. Prediction of Conflicts in Genetic Variant Classifications Through Machine Learning and Analysis of Predictive Features. 1, 1 (November 2018), 20 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Ever since Gregor Mendel established the fundamental relationship between genes and visible morphological and physiological characteristics, the field of genetics has grown exponentially [6]. Because these genes contain the instructions necessary for all living organisms, changes made to them can have detrimental effects on the organism and are a major public health concern. Some examples of prominent genetic variants causing human disease include GAA gene variants, which are responsible for causing Pompe disease [2]; DMD gene variants, which are responsible for causing Duchenne muscular dystrophy (DMD) [12]; and GALNS and GLB1 gene variants, which are responsible for causing Type IV Mucopolysaccharidosis [10]. However, not all these genetic variants are harmful. For example, a single nucleotide substitution can result in production of an identical protein.

Thus, one of the primary challenges in genetics is classification of the harmfulness of a genetic variant. Currently, The American College of Genetics and Genomics (AMCG) has guidelines for the standardization of classification of genetic variants into a total of five categories: pathogenic, likely pathogenic, variant of unknown significance (VUS), likely benign, and benign [9]. Although these guidelines exist, there are discrepancies in the classification done by different clinical laboratories as different laboratories have access to different phenotypic evidence, interpretations of the literature on specific variants, and history-weighting algorithms [3]. These conflicting classifications become problematic when a clinician orders a test, receives the result, checks the result against information available in free databases, and sees conflicting results. At its most extreme, the difference between a VUS or pathogenic variant classification could either result in an urgent surgery or no treatment at all. In this extreme case, we can see the importance of accurate and unified genetic classifications.

With this project, we set out to accomplish two goals: (1) predict conflicting classifications with high accuracy, sensitivity, specificity, and AUROC, and (2) identify the features that are most relevant in the model's determination of conflicting or no conflicting classifications. We will ensure that our model not only is able to classify variants appropriately with high accuracy, but also will be explainable with clear identification of the features which have the largest impact on classification. This explainable AI model will be relevant for clinical laboratories, medical researchers, and clinicians, who can use our model to predict conflicts and also be aware of which features have large bearings on whether or not a variant will have conflicting classifications.

2 LITERATURE REVIEW

Each year more than a million people are affected by genetic disorders as the human body has around 20,000 - 25,000 genes and any slight alteration to a single gene can lead to development of genetic variants, causing harmful genetic diseases in some cases [?]. Approaches to tackling this major issue are being revolutionized by the widespread use of next-generation sequencing to rapidly sequence large amounts of DNA. This workflow makes it significantly easier to detect genetic variants [5]. In response to this rapid influx of data, researchers at the National Center for Biotechnology Information (NCBI) has created a comprehensive archive known as ClinVar where researchers and clinical laboratories can submit their interpretation of the clinical significance of genetic variants [5]. However, the major issue with this open submission archive format is that classification of a variant into one of the five AMCG categories requires the consideration of numerous factors including interpretation of literature, functional studies, and phenotype evidence – and access to these pieces of information vary widely. The differences in these factors across submitting parties is a major cause in conflicting classifications of genetic variants.

In the domain of cancer variants, researchers have discovered that conflicting classifications occur quite frequently between clinical labs and ClinVar [3] as well as amongst clinical labs [1]. Gradishar et al. found that of the 4,250 BRCA1 and BRCA2 variants, 73.2% had fully concordant classifications, 12.3% had partially concordant classifications, and 14.5% had discordant classifications. In this case, 14.0% of the discordant classifications occurred when the clinical laboratory assigned a definitive classification of either pathogenic or benign and ClinVar has an uncertain classification [3]. The source of the differing classifications appears to be the laboratories' phenotypic evidence from a history-weighting algorithm that allows it to give a definitive diagnosis [3]. In the opposite case (0.3%), where the database gives a definitive classification and the laboratory gives an uncertain classification, ClinVar uses the +1 canonical splice site as strong evidence of pathogenicity, but there have been cases of individuals with an alternative transcript that essentially negates cancer risk with enough functional protein [3]. Thus, ClinVar's general rule of thumb conflicts with the laboratory's deferral to a lack of clinical evidence [3]. Lastly, opposite classifications (benign vs. pathogenic or vice versa) make up only 0.1% of all classifications. In one of these cases, ClinVar's use of the +1 splice site resulted in a pathogenic classification, which disagreed with the literature and phenotypic evidence used by the clinical laboratory to classify the variant as benign [3]. In sum, although there appear to be an alarmingly large percentage of conflicting classifications, the vast majority of them are due to uncertain classifications by ClinVar [3]. The reasons for these conflicts seem to be due to rules of thumb, literature, and phenotypic evidence.

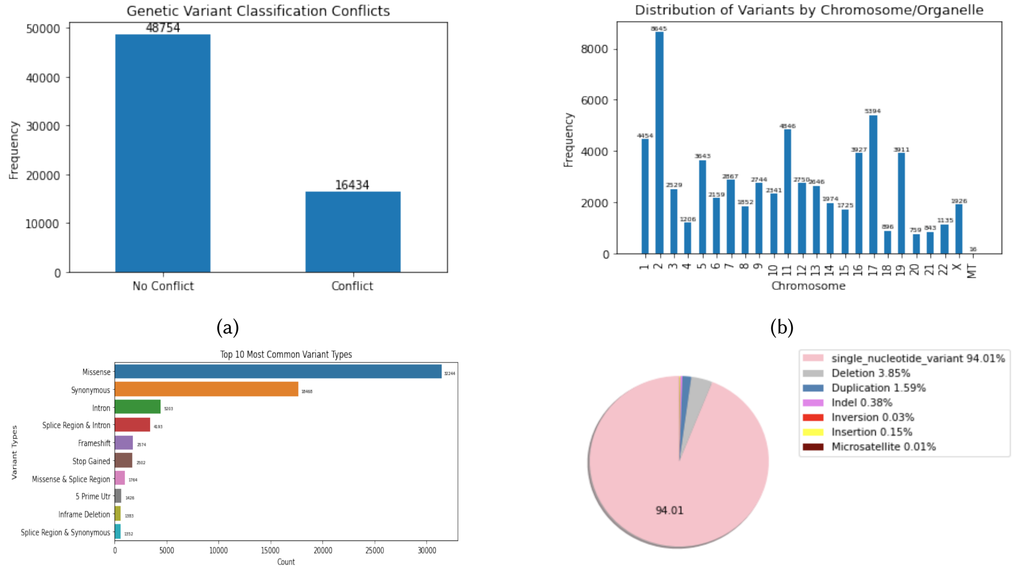
Because the issue of conflicting classifications in ClinVar is well known, collaborative efforts have already been made to decrease the number of conflicting classifications [4]. Four clinical laboratories which submit their variants to ClinVar collaborated to resolve differences in classification and were able to increase concordant classifications from 88.3% to 91.7% by reassessing current criteria and/or internal data sharing [4].

Although much research has been done in the adjacent space of applying machine learning methods to classify genetic variants into the five AMCG categories [7] or as causing certain diseases [Subhani and Anjum], only two papers have been published that directly address the issue of conflicting classifications of variants [6, 11]. In one study, the authors used logistic regression, decision trees, random forest, gradient boost classifier, and neural networks [11] and found that the random forest classifier performed the best with 0.76 precision, 0.96 recall, and 0.86 F1 score. However, all their models struggled to classify conflicts. In another study, the authors used both Random Forest and Gradient Boosting classifiers and found that the two models performed nearly identically with regards to recall, precision and F1 score [6]. However, classification time was one order of magnitude faster for the Gradient Boosting classifier compared to the Random Forest classifier [6]. With this in mind, the authors advocated for the Gradient Boosting classifier and were able to achieve 77% accuracy in classification. There are two major gaps in understanding in the literature regarding the classification problem. The first is the careful preservation of the biological meaning in data when converting it into a numerical format and transparency surrounding how this is done. The second is the lack of understanding how these machine learning models come to a decision – essentially unraveling the blackbox of machine learning. Our paper aims to address these two aims by explaining how we preserve the biological meaning of our data in our data pre-processing steps and thoroughly analyzing the features which are most relevant in the binary classification of genetic conflicts.

3 METHODS

3.1 Data Description

The original dataset used for this study is from ClinVar's NCBI [5]. The actual dataset used for our study is a filtered version of the original dataset designed to specifically focus on genetic variants with conflicting classifications taken from Kaggle (https://www.kaggle.com/datasets/kevinarvai/clinvar-conflicting). Conflicting classifications were defined as when two of any of the three following categories were present for one variant: (1) likely benign or benign, (2) VUS, or (3) likely pathogenic or pathogenic. Because this dataset focused specifically on conflicting classifications, all variants in the original ClinVar .vcf file containing only a single classification were omitted from the dataset. After this initial filtering, the dataset's dimensions are as follows: 65,187 genetic variants with a total of 46 features for each variant. A comprehensive list of the biological relevance of all 46 features will not be included because many features were dropped and a comprehensive list of the relevant features is shown in Table 1. The relevant feature that we aim to predict is 'CLASS' with 0's denoting no conflicting classifications and 1's denoting conflicting classifications. In our initial exploratory data analysis, we found that there is a severe data imbalance between classes as depicted in Figure 1a with approximately 75% of the data not having any conflicts and approximately 25% of the data having conflicts.



(a) (b)

(c) (d)

Fig. 1. Analysis of the ClinVar Dataset (a) bar plot of the number of genetic variants which have no conflicts or have conflicts (b) bar plot of the of the number of variants on within chromosome/organelle (c) bar plot of the distribution of the top ten genetic variant types (d) pie chart showing the distribution of the type of mutation

Regarding the distribution of the genetic variants across chromosomes, we found that they are distributed across all 22 autosomal chromosomes, the X chromosome, and in the mitochondria (MT) with variants being most commonly found on autosomal chromosome 2 as shown in Figure 1b. Additionally, we can see that there are no variants on the Y chromosome, meaning that Y chromosomal variants most likely only have a single classification. In Figures 1c and 1d, we can see that the vast majority of variants are missense or synonymous mutations that result from the alteration of just a single nucleotide (single_nucleotide_variant).

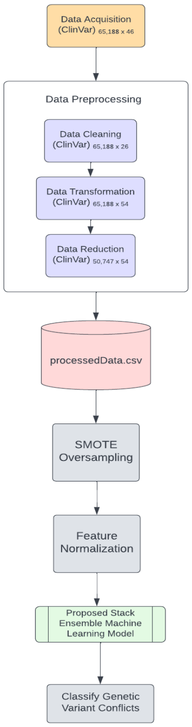


Fig. 2. Flowsheet of Data Processing

3.2 Data Cleaning

Features including 'MOTIF_SCORE_CHANGE', 'HIGH_INF_POS', 'MOTIF_POS', 'MOTIF_NAME', 'DISTANCE', 'SSR', 'CLNSIGINCL', 'CLNDNINCL', and 'CLNDISDBINCL' were missing more than 90% of their data. After dropping these features from the dataframe, we reduced the feature set from 46 to 37. Among these features, several of them do not contain biologically relevant information. These features include 'CLNDISB' (tag-value pair of the database and the variant identifier), 'CLNVI' (tag-value pairs of the database and the variant identifier), 'Feature' (an ensemble stable ID), and 'BAM_EDIT' (ability to edit using a BAM file is not biologically relevant). Additionally, other features contain biological information that are already captured in other features such as 'Allele' (identical to 'ALT'), 'Consequence' (identical to MC, but lacking comorbidity), 'Codons' (information contained within 'REF', 'ALT', and 'Amino_acid'), and 'CLNHGVS' (information contained within 'REF' and 'ALT') . The features named 'BIOTYPE' and 'Feature_type' did not contribute any discriminatory power as all the data values were simply 'transcript' and 'protein_coding'. Lastly, the 'BLOSUM62' feature was also omitted because imputing scores from the 'BLOSUM62' matrix is only possible with genetic variants in which a single amino acid is converted to another amino acid. However, we also wanted to include genetic transformations with multiple amino acids to one, or one to many. The biological significance of the final 26 initial features are described in Table 1.

Feature Name	Biological Relevance
CHROM	Genetic location of a variant (all autosomal chromosomes, X, and MT for mitochondrial)
POS	Position of the variant on the chromosome
REF	Reference Allele
ALT	Alternate Allele
AF_ESP	Allele frequency from GO-ESP
AF_EXAC	Allele frequency from ExAC, which is now known as gnomAD
AF_TGP	Allele frequency from the 1000 Genomes Project
CLNDN	ClinVar's preferred disease name for the concept specified by disease identifiers in CLNDISDB
CLNVC	Variant Type (i.e. single nucleotide variant, deletion, insertion, indel etc.)
MC	Molecular Consequence(s) (i.e. missense variant, synonymous variant, etc.)
ORIGIN	The allele origin (i.e. germline, somatic, inherited, paternal, maternal etc.)
CLASS	**Conflicting (1) or no conflicting classifications (0)**
IMPACT	Rating of the impact of the variant (low, moderate, high, and modifier)
SYMBOL	Gene Name
EXON	Exon #/Total # of Exons
INTRON	Intron #/Total # of Exons
cDNA_position	Relative position of base pair in cDNA sequence
CDS_position	Relative position of base pair in coding sequence
Protein_position	Relative position of amino acid in protein
Amino_Acids	Original/Altered Amino Acid (i.e. E/D)
Strand	Forward (+1) and Reverse (-1)
SIFT	the SIFT prediction and/or score, with both given as prediction(score)
PolyPhen	the PolyPhen prediction and/or score
LoFtool	Loss of Function tolerance score for loss of function variants
CADD_PHRED	Phred-scaled CADD score
CADD_RAW	Score of the deleteriousness of variants: http://cadd.gs.washington.edu/

Table 1. Features in ClinVar and their Biological Relevance.

3.3 Data Transformation

Many of the features in this dataset are non-numerical, requiring conversion of these variables into numerical/dummy variables. A brief description of how each of these features were processed are detailed in Table 2 below. In some cases, the original feature column was dropped and additional columns were appended to the data frame (indicated as 'original feature name' → 'replaced features').

Feature Name	Pre-Processing
CHROM	The 22 autosomal chromosomes were left as integers and variants on the X chromosome were assigned to 23 and mitochondrial DNA variants were assigned to 24.
REF	The four base pairs A, T, G, and C were assigned the dummy variables 1, 2, 3, and 4. Cases of non single nucleotide variants, where there were more than a single nucleotide in the REF column, were assigned to the dummy variable 5.
ALT	In the same way that dummy variable assignment occurred for REF, the four base pairs A, T, G, and C were assigned the dummy variables 1, 2, 3, and 4. Similarly, cases of non-single nucleotide variants were assigned to the dummy variable 5.
CLNVC	This specific feature consists of seven unique strings: ‘single_nucleotide_variant’, ‘Deletion’, ‘Duplication’, ‘Indel’, ‘Inversion’, ‘Insertion’, and ‘Microsatellite’, which were then assigned dummy variables from 0-6.
MC ⟶ (appendix 1a)	Because this feature was formatted as a comma-separated list of “disease identifier|type of variant”, we chose to convert MC into one-hot-encoded rows for missense variants, synonymous variants, and other variants.
ORIGIN ⟶ (appendix 1b)	Origin is formatted in such a way that one or more of the following values correspond to various different categories: 0 - unknown; 1 - germline; 2 - somatic; 4 - inherited; 8 - paternal; 16 - maternal; 32 - de-novo; 64 - biparental; 128 - uniparental; 256 - not-tested; 512 - tested-inconclusive; 1073741824 - other. Thus, a variant with a 49 in the ORIGIN feature would correspond to being de-novo (32), maternal (16), and germline (1) in origin. After parsing through these numbers, we also one-hot encoded this feature.
CLNDN ⟶ (appendix 1c)	This feature is a pipe-separated list of disease names that could be caused by the genetic mutation. We separated out the list and created dummy variables through one-hot encoding for the top 15 diseases, excluding ‘not_specified’ and ‘not_found’.
IMPACT	This feature consists of four unique strings: ‘MODERATE’, ‘MODIFIER’, ‘LOW’, ‘HIGH’, which were assigned the dummy variables 0-3.
Amino_acids ⟶ (appendix 1d)	This feature was formatted as a string in the following manner: ‘original amino acid(s)/altered amino acid(s). We chose to assign numbers to the one letter amino acid codes in alphabetical order (A for Arginine being denoted as 1, C for Cysteine being denoted as 2, etc.).
SIFT	SIFT(sorting intolerant from tolerant) values are assigned values as follows: tolerated: 1, deleterious_low_confidence: 2, deleterious : 3
PolyPhen	PolyPhen(polymorphism phenotyping) values are assigned values as follows : benign: 1, probably_damaging: 2, possibly_damaging : 3
SYMBOL	This feature contains 2329 unique gene names, which were arbitrarily assigned dummy variables from 0-2328.

INTRON/EXON → (appendix 1e)	This feature contained strings of the 'Exon # / Total # Exons' or NaN's. Because a variant can only occur on an intron or an exon, a variant can have an 'INTRON' as NaN and 'EXON' is '1/5'. Because of this, INTRON and EXON were replaced with IE and IE_Loc with the former indicating the variant's presence on an intron (0) or exon (1) and the latter corresponding the variant's location fractional float.

Table 2. Pre-Processing of Non-Numerical Features

3.4 Data Reduction

After completing our data pre-processing portion, our processed dataframe had 54 feature columns containing purely numerical data. From there, we needed to drop any genetic variants with missing values in any of the features as these could not be inputted into a machine learning model. After this was done, the data frame contained 50,747 variants with 54 features.

3.5 Oversampling and Feature Normalization

Due to a large data imbalance where the vast majority of the data had non-conflicting genetic variant classifications, we used Synthetic Minority Oversampling TEchniques (SMOTE) to combat this issue and equalize the number of variants of 'CLASS' 0 and 1, which is depicted via t-SNE plots in Figure 3.

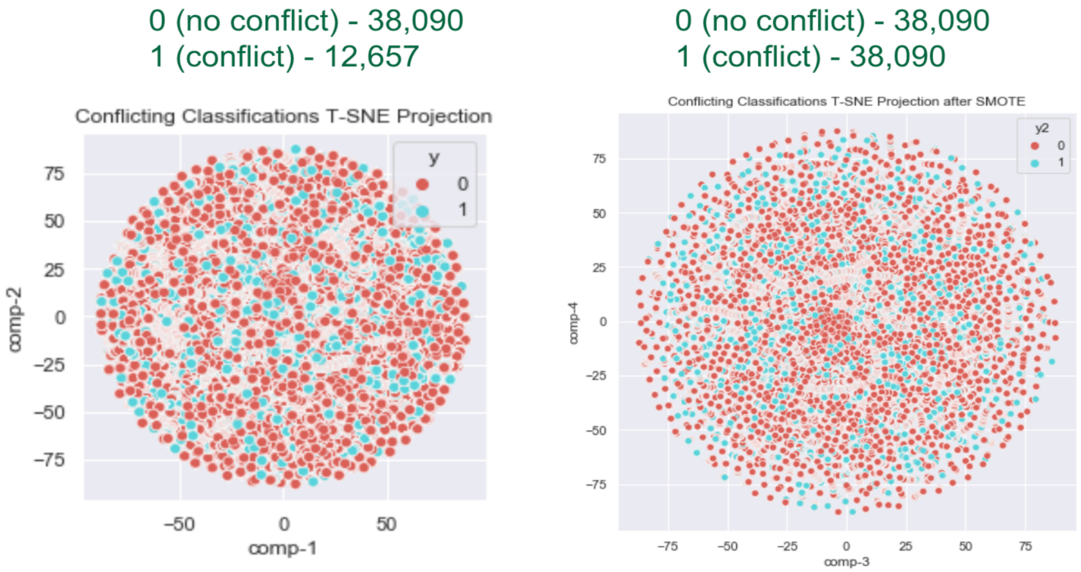


Fig. 3. t-SNE Projection Plots of the dataset before and after SMOTE

Another issue we needed to combat were the large differences in magnitude between features. For example, values in 'POS' were as large in magnitude as 2.47 x 108, while other features such as 'germline' were one-hot-encoded. Thus, in order to prevent the features with larger magnitudes from dominating the models, we normalized the features to have a mean of 0 and a standard deviation of 1.

3.6 Machine Learning Models

We chose a total of 10 different binary classification models to tackle this problem: Random Forest, SVM, XGBoost, K-nearest neighbors, Logistic Regression, Decision Tree Classifier, AdaBoost, Gradient Boost, Gaussian Naive Bayes, and lastly a Stacked Ensemble Machine Learning Model. We used this large number of different machine learning models to simply get a sense of which models might perform well. For each of the first nine listed models, we used an 70/30 train-test split and used the default hyperparameters found in the sci-kit learn library. The reasoning behind this choice for hyperparameters was simply to get a sense of which models performed the best and could then be further tuned and be inputted into our stacked ensemble model. Because the random forest, SVM, and XGBoost models performed the best initially, we then inputted them into an initial stacked ensemble model as depicted in Figure 4. The results of these initial untuned models are shown in Table 3, which can be found in the Results section. Because the stacked ensemble model outperformed individual Random Forest, XGBoost, and SVM models, we decided to focus on these base models and tune their hyperparameters. Turning was done based on the accuracy metric of these three models using a randomized search cross validation method. Unfortunately, hyperparameter tuning for SVM was not possible due to time constraints as there were issues with using the randomized search cross validation method and grid-search needed to be used. The results from the tuned models are shown in Table 4.

3.6.1 Stacked Ensemble Machine Learning Model. To create the best model for classifying consistent and conflicting classification of genetic variants, we proposed the use of a stacked ensemble machine learning model. This model is an ensemble learning technique that amalgamates a number of base-level machine learning models in a particular configuration along with a meta-learning classifier to address our genetic variant classification problem. Our pre-processed data is used for training the base level machine learning models and outputs from these models are then used as input for training the meta-learner and getting the corresponding prediction for genetic variant classifications. The steps involved in training the stacked ensemble machine learning model are as follows:

(1) The preprocessed dataset of genetic variant classification is split into 70% training and 30% testing.
(2) Three base-level machine learning models (Random Forest, SVM, and XGBoost) with appropriate hyperparameter tuning for XGBoost and Random Forest base models are trained along with a standard Logistic Regression as a meta-learner classifier.
(3) Predictions are obtained from the baseline machine learning models and are fed as input to the meta-learner classifier to predict our binary conflicting classifications.

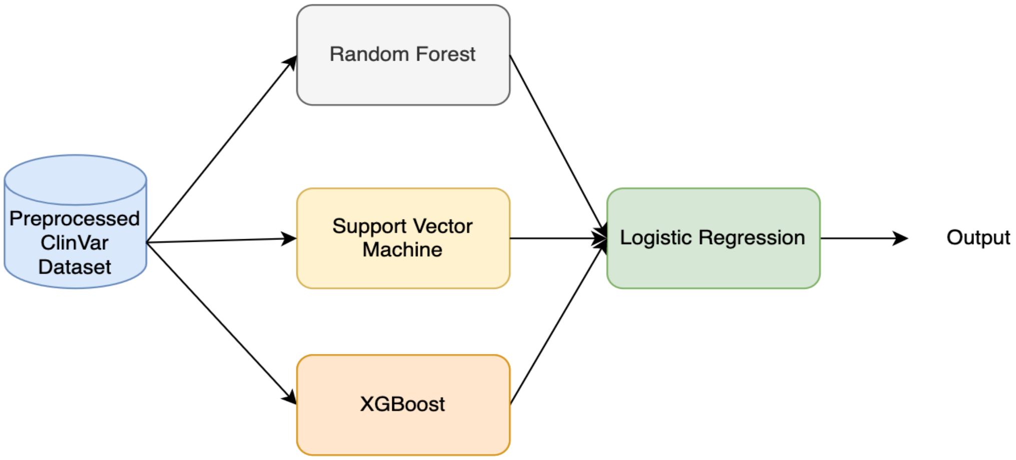


Fig. 4. Stacked Ensemble Model Flow

3.7 Evaluation Metrics

For our evaluation metrics, we chose to look specifically at accuracy, sensitivity, specificity, and AUROC. The accuracy metrics captures the performance of the models across both classes and thus was important to include. Sensitivity and specificity are also important metrics that capture information not captured in the accuracy metric. Sensitivity captures the ability of a model to classify true positives as true positives and specificity captures the ability of a model to classify true negatives as true negatives. Additionally, these two metrics typically act in competing fashions where improvements in one metric result in decreased performance in the other and thus including both gives the best understanding of the performance of our models. Lastly, AUROC is an additional important metric in determining how capable a model is at distinguishing classes at various different thresholds.

4 RESULTS

As briefly mentioned earlier, Table 3 contains the performance of the 9 initial models with the default hyperparameters. Additionally, it also includes the performance of the initial stacked ensemble model using the untuned hyperparameters from three high-performing models (Random Forest, SVM, and XGBoost).

Model Name	Sensitivity	Specificity	Accuracy	AUROC
Random Forest	0.78	0.87	0.83	0.91
SVM	0.73	0.77	0.75	0.84
XGBoost	0.8	0.87	0.84	0.92
K-nearest neighbors	0.69	0.65	0.67	0.72
Logistic Regression	0	1.0	0.5	0.51

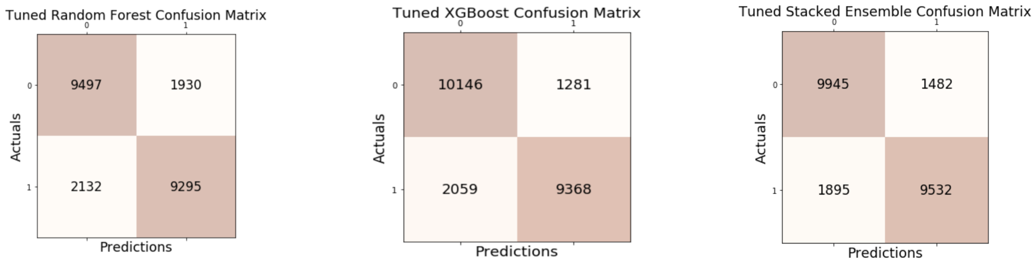
Decision Tree Classifier	0.79	0.77	0.78	0.78
AdaBoost	0.78	0.78	0.78	0.87
Gradient Boost	0.76	0.79	0.78	0.87
Gaussian Naive Bayes	0	1.0	0.5	0.53
Stacked Ensemble	**0.84**	**0.86**	**0.85**	**0.93**

Table 3. Initial Metrics of Machine Learning Models Against Proposed Stacked Ensemble Learning Model

Table 4 contains the performance of the two of the three top performing models after hyperparameter tuning as well as the stacked ensemble model after each of the respective constitutive models' hyperparameters were tuned. The confusion matrices for the three models as well as their corresponding AUROC curves are shown below in Figure 5 and Figure 6.

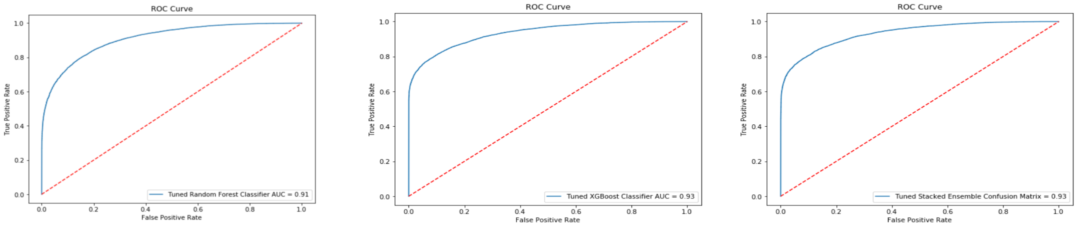
Model Name	**Sensitivity**	**Specificity**	**Accuracy**	**AUROC**
Random Forest	0.81	0.83	0.82	0.91
XGBoost	0.82	0.89	0.85	0.93
Stacked Ensemble	**0.83**	**0.87**	**0.85**	**0.93**

Table 4. Performance of Models After Hyperparameter Tuning



(a) (b) (c)

Fig. 5. Tuned Confusion Matrices for 3 models: (a) Random Forest (b) XGBoost **(c)* Proposed Stacked Ensemble**



(a) (b) (c)

Fig. 6. Tuned AUROC for 3 models: (a) Random Forest (b) XGBoost **(c)* Proposed Stacked Ensemble**

4.1 Analysis of Salient Features

With these results shown in Table 4, we will now address our second aim and decode the intricacies behind our models through the use of several different tools.

4.1.1 LIME - Local Interpretable Model-Agnostic Explanations.

Local Interpretable Model-Agnostic Explanation (LIME) is a tool for decoding the intricacies of black box machine learning models and uses local surrogate models instead of global surrogate models to generate individual predictions. LIME has a local aspect for describing individual predictions, is model-agnostic, and is able to train with any kind of machine learning models like decision tree or support vector machine. Additionally, LIME generates a new dataset by perturbing the present dataset and provides individual predictions from local surrogate models for any given variation of data for a particular machine learning model. On new datasets, LIME trains an interpretable model that calculates the closeness of a sampled instance to an instance of interest. For the problem of genetic variant conflict classification, LIME can provide us with important information regarding which features have a significant effect on model prediction. Our LIME trains local surrogate models and follows the given approach with the results depicted in Figure 7:

1. Select several features like missense_variant, synonymous variant, and others for the stacked ensemble model as instances of interest for decoding the black box model prediction and explaining individual feature contribution for the particular target class
2. Perturb the dataset
3. Train the model on new dataset points
4. Weight new samples in accordance with closeness to the instance of interest
5. Train the model again on the perturbed dataset
6. Derive results by interpreting the prediction of the local model

As we can see from Figure 7, the ‘missense_variant’ feature is a contributor towards consistent classification. On the other hand, the ‘synonymous_variant’, ‘has_Hereditary_cancer-predisposing_syndrome’, and ‘other_Variant’ features are contributors towards conflicting classifications.

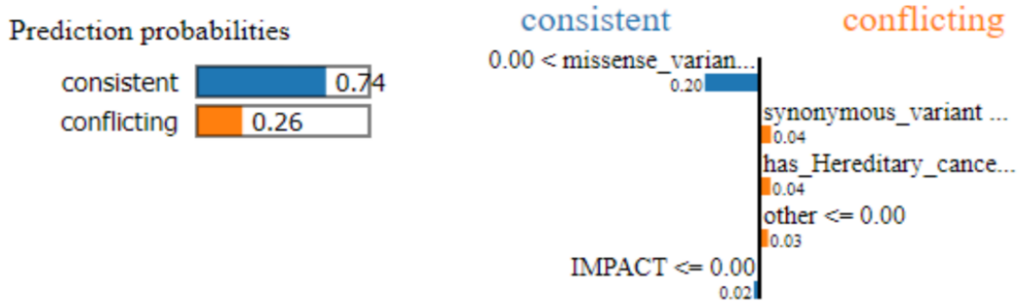


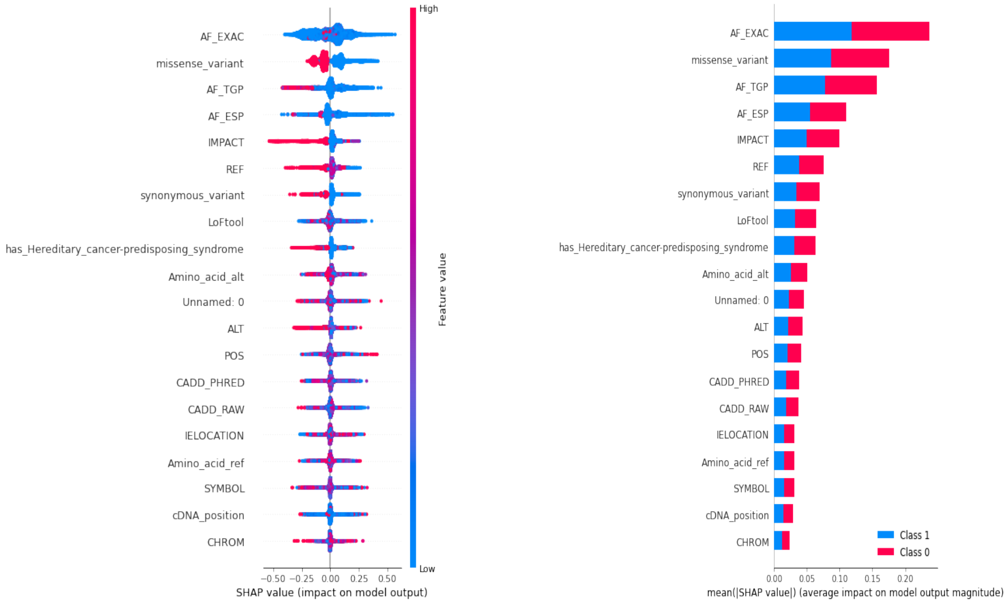
Fig. 7. LIME Feature Salience

4.1.2 SHAP - Shapley Additive Explanations.

Shapley Additive Explanations (SHAP) have numerous visualization tools that are useful for interpreting the contribution of each feature value in predicting whether or not a variant will have a conflicting classification. It functions as a form of coalitional game theory where Shapley values indicate the distribution of payout (prediction) among the features. There are a wide array of SHAP visualization tools, but here we will visualize just two of them: the SHAP summary plot with variable importance (Figure 6a) and the SHAP Summary Plot by class label (Figure 6b) to develop an intuition and understanding for our model's decision-making.

a Figure 8a shows dots representing the feature value of conflicting genetic variants for each individual data instance. Red dots represent a high value of contribution towards conflicting genetic variant classes and blue dots represent low values of contribution. Along with the color of the dots, their location on the x-axis determines their contribution towards predicted outcomes. Dots with SHAP values greater than 0 contribute positively while those with SHAP values less than 0 contribute negatively. From this plot, we can see that AF_EXAC has the most predictive power among all features.

b Figure 8b lists all features in order of most significant to least significant. Almost all features illustrate the pattern of equal contribution towards being categorized as a genetic variant with conflicting classifications (label=1) or consistent classifications (label = 0) as each rectangle representing each feature has approximately equal areas of red and blue.



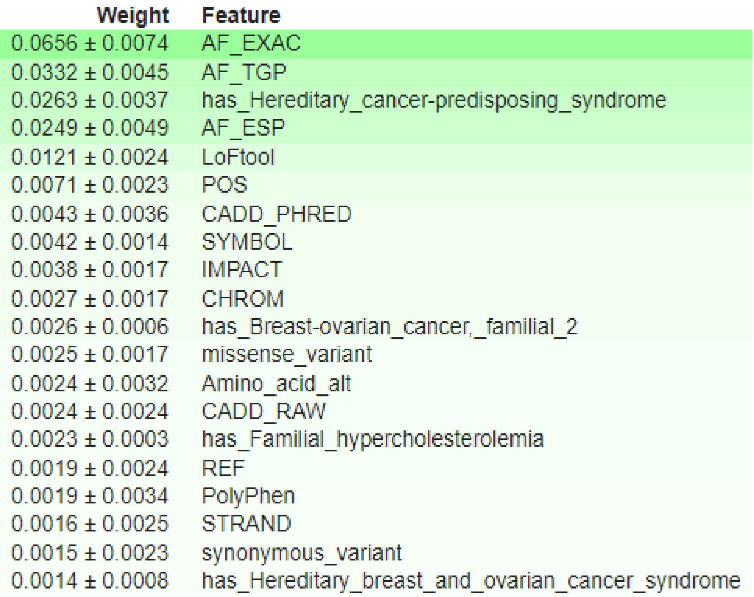
(a) (b)

Fig. 8. SHAP Summary Plot by (a) variable importance (b) class

4.1.3 Permutation Feature Importance.

Permutation feature importance performs a similar role to SHAP and acts as another independent assessment of which features contribute most to our models' ultimate classification of a variant having conflicting vs. no conflicting classifications. Permutation feature importance works by randomly shuffling the values for a particular feature and then calculating the performance of the machine learning model. For example, in our case if we shuffle feature values for 'AF_EXAC' feature randomly, we can determine the effect of the 'AF_EXAC' feature on model performance. If there is a significant depreciation in the performance of our machine learning model due to permuting feature values for 'AF_EXAC', this signifies that the information in 'AF_EXAC' was crucial in determining the final machine learning model outcome. In the reverse scenario, if we see a negligible decrease in performance of the machine learning model, this demonstrates that the shuffled feature did not have a significant impact on the model's performance. The steps involved in permutation feature importance are as follows:

1. Shuffle feature values of one feature and keep other features constant
2. Train machine learning algorithms using datasets with shuffled feature values for evaluating prediction outcomes relative to ground truth labels
3. Calculate feature importance score for a particular feature by evaluating the depreciation in the model's performance
4. Rank all features according to their feature importance scores in tabular format in descending order

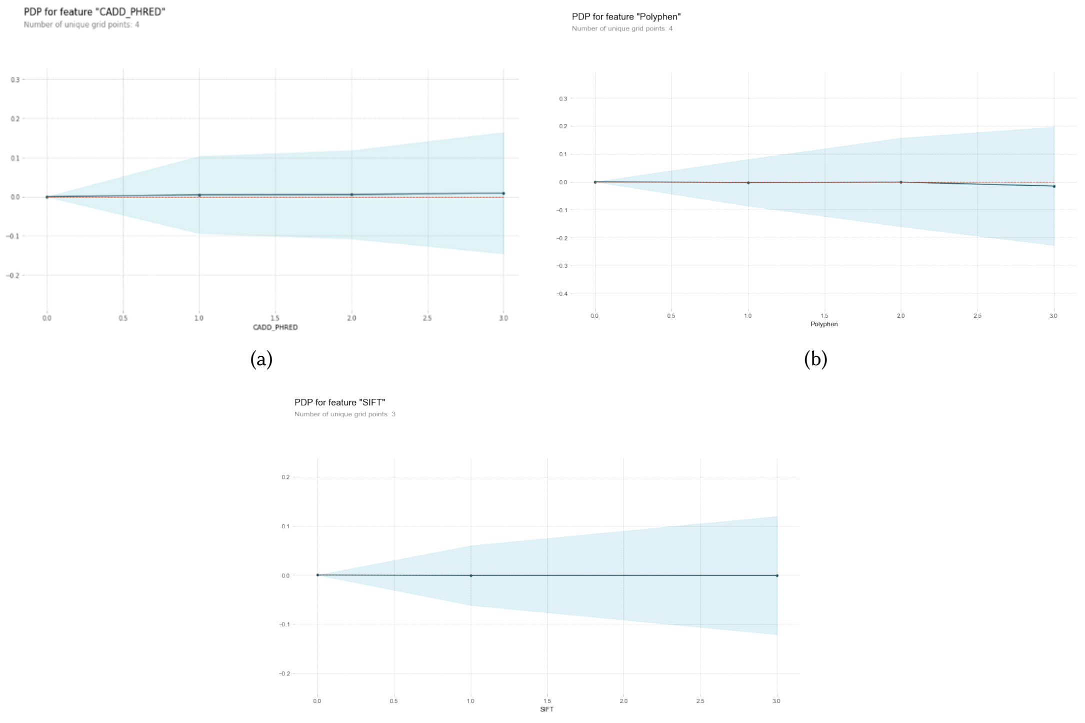


Weight	Feature
0.0656 ± 0.0074	AF_EXAC
0.0332 ± 0.0045	AF_TGP
0.0263 ± 0.0037	has_Hereditary_cancer-predisposing_syndrome
0.0249 ± 0.0049	AF_ESP
0.0121 ± 0.0024	LoFtool
0.0071 ± 0.0023	POS
0.0043 ± 0.0036	CADD_PHRED
0.0042 ± 0.0014	SYMBOL
0.0038 ± 0.0017	IMPACT
0.0027 ± 0.0017	CHROM
0.0026 ± 0.0006	has_Breast-ovarian_cancer,_familial_2
0.0025 ± 0.0017	missense_variant
0.0024 ± 0.0032	Amino_acid_alt
0.0024 ± 0.0024	CADD_RAW
0.0023 ± 0.0003	has_Familial_hypercholesterolemia
0.0019 ± 0.0024	REF
0.0019 ± 0.0034	PolyPhen
0.0016 ± 0.0025	STRAND
0.0015 ± 0.0023	synonymous_variant
0.0014 ± 0.0008	has_Hereditary_breast_and_ovarian_cancer_syndrome

Fig. 9. Permutation Feature Importance

4.1.4 *Partial Dependence Plots.*

A partial dependence plot (PDP) provides visualization of the marginal effects of predictor variables like PolyPhen, SIFT, and CADD_PHRED on model performance by plotting the average model outcome for numerous predictor values. The PDPs use a partial dependence function to illustrate predictive probabilities for conflicting classification against different input values.



(a)

(b)

(c)

Fig. 10. Partial Dependency Plots for (a) CADD_PHRED (b) PolyPhen (c) SIFT

5 DISCUSSION

After feature normalization and SMOTE, we saw a huge improvement in our models, especially with the sensitivity metric. As seen in Table 3, the Random Forest, SVM, XGBoost, Decision Tree Classifier, AdaBoost, and Gradient Boost models all performed well across all metrics. The KNN model performed less well across all metrics, but still at an adequate level. In sharp contrast to all these models, the Logistic Regression and Gaussian Naive Bayes models performed quite poorly – simply assigning all variants to the conflicting classification class. After tuning the hyperparameters of the Random Forest and XGBoost base models and inputting these tuned base models into the stacked ensemble model, we saw essentially no improvement in the sensitivity, specificity, accuracy, or AUROC, which was quite surprising. Perhaps this is because our models have already achieved the upper threshold of performance and further hyperparameter tuning simply results in overfitting of the training data. Additionally, we found that the 'AF_EXAC' feature was universally rated as having the largest effect on the classification of genetic variants.

There are several limitations to our study that could be improved upon in future works. First, in our assignment of features to categorical variables, we opted to assign variables to certain numbers (i.e. for 'REF' and 'ALT' nucleotide features, A ⟶ 1, T ⟶ 2, G ⟶ 3, and C ⟶ 4 as well as 'CHROM' where the X chromosome ⟶ 23 and mtDNA variants ⟶ 24) when it would have made more sense to one-hot encode such features in order to avoid implying a continuous relationship between the categorical variables. Perhaps if this was done, we would see even better results. In a similar vein, we also struggled to convert genetic variants that were not single nucleotide variants into a biologically meaningful numerical format. For example, in the case of a hypothetical deletion mutation (i.e. A ⟶ ATGCC), we assigned 'REF' to 1, but assigned 'ALT' to 5 and lumped cases when 'REF' or 'ALT' were greater than 1 into the numerical category of 5. As a result, our models had no way of distinguishing between different types of non-single nucleotide variants. Additionally, another limitation in our data was the several potentially features that needed to be dropped due to the large amounts of missing data. However, it remains a possibility that these gaps in data will be filled as more information on variants is discovered and added to ClinVar. Another important limitation of our study is that our results only represent a single model's run and we did not implement the standard 10-fold cross validation to ensure that good model performance was not simply due to luck in the train-test split. Our reasoning is that our dataset was large enough that 10-fold cross validation was not necessary. Lastly, we did minimal hyperparameter tuning – only tuning the hyperparameters for the Random Forest and XGBoost models. It would be important to tune hyperparameters for all nine of our baseline models to determine which ones truly perform the best and input those models into the stacked ensemble model.

With sufficient time, we plan to pursue solutions to these limitations of our study, especially regarding cross validation and further hyperparameter tuning of our baseline models. In the future, it would be interesting to conduct this same research on the most up-to-date ClinVar dataset as there will be even more variants and more fleshed out features. Additionally, it would be interesting to see how this work can be applied to other issues of conflicting classifications in the medical realm.

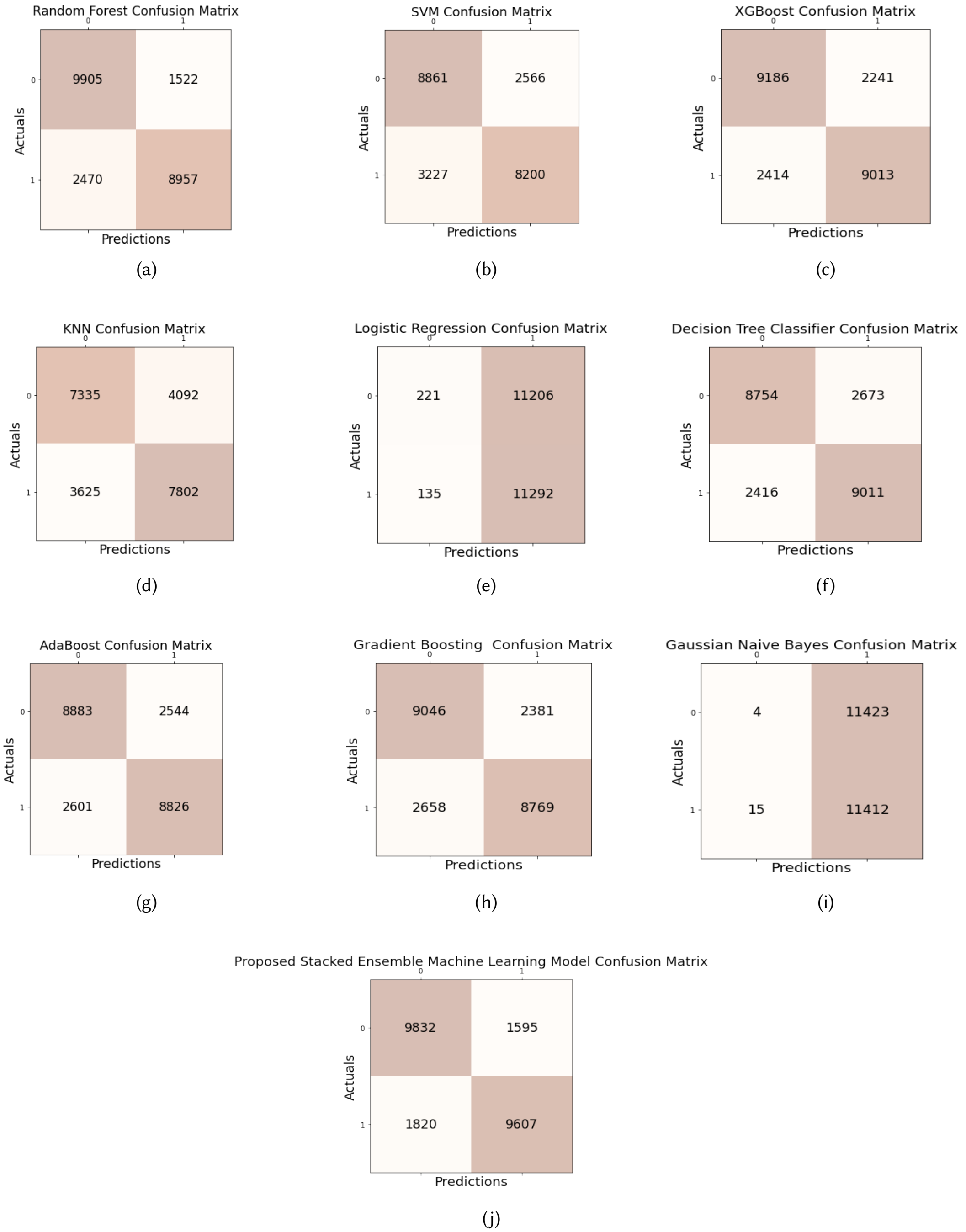
6 CONCLUSION

With this project, we detected conflicting genetic variants from the medical research dataset ClinVar using the proposed stacked ensemble machine learning model and achieved 85% accuracy, 83% sensitivity, and 87% specificity along with a significant AUROC of 0.93. Additionally, to decode predictions made by the black box machine learning model, we used various interpretable machine learning techniques including LIME, SHAP, Permutation Feature Importance, and Partial Dependence Plots to help healthcare practitioners understand the reasoning behind our particular machine learning model's prediction. Our proposed stacked ensemble machine learning model will foster medical research in the domain of genetics by handling the excessive conflicts in genetic variant classification. Our model's predictions can allow healthcare practitioners to focus on patient care rather than agonizing over the pathogenicity of a genetic variant. The work can potentially be extended to other medical datasets to resolve conflicting classification concerns.

7 ACKNOWLEDGEMENTS

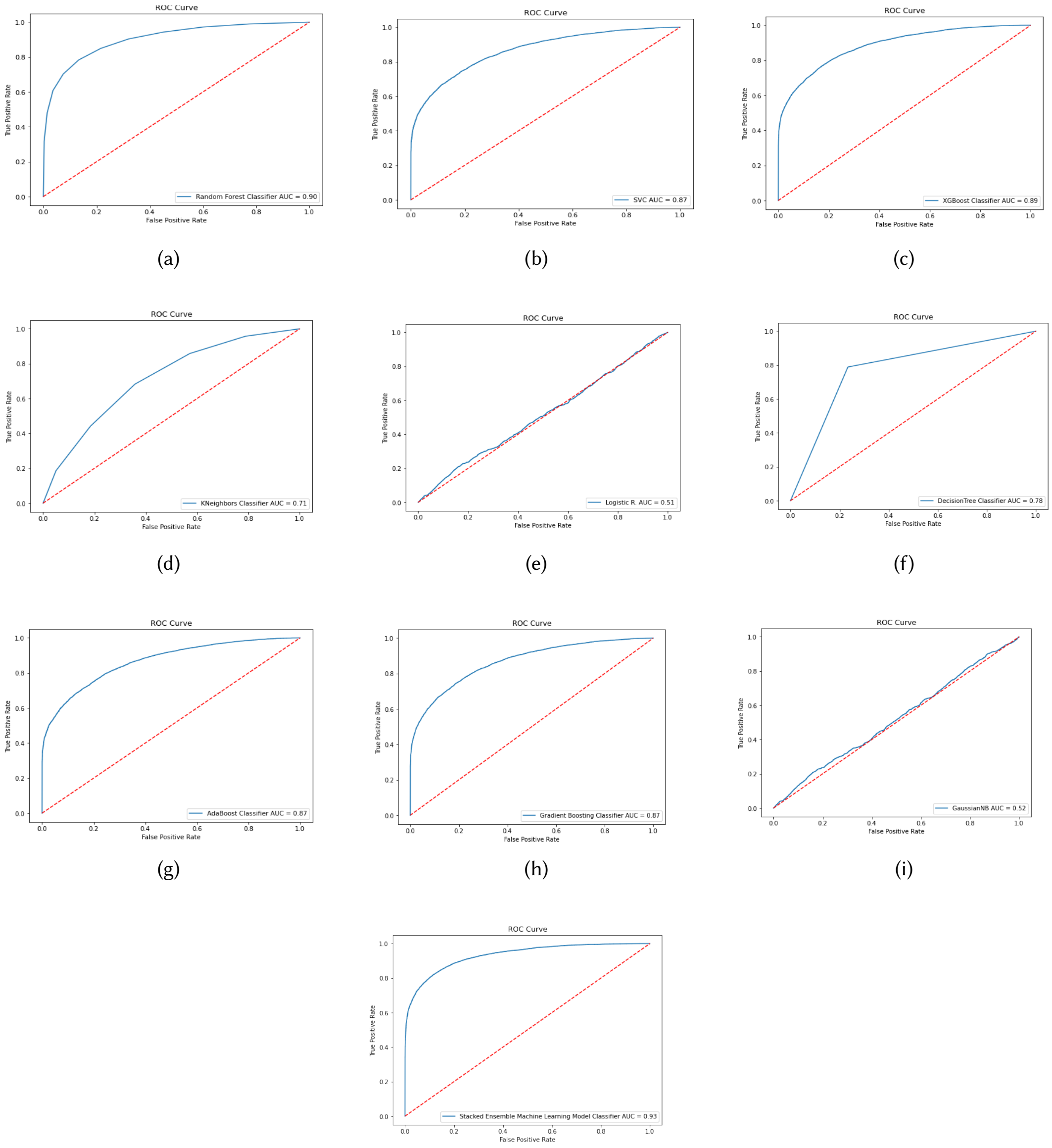
This research project was part of the Data Science for Health course in Fall 2022. Thank you to our Dartmouth professor Temiloluwa Prioleau and our TAs Prajakta Belsare and Wayner Barrios for their feedback and guidance.

SUPPLEMENTARY FIGURES



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j)

Fig. 11. Confusion Matrices for 10 different models: (a) Random Forest (b) SVM (c) XGBoost (d) K-Nearest-Neighbor (e) Logistic Regression, (f) Decision Tree Classifier, (g) AdaBoost, (h) Gradient Boost, (i) Guassian Naive Bayes, **(j)*** **Proposed Stacked Ensemble**



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j)

Fig. 12. AUROC for 10 different models: (a) Random Forest (b) SVM (c) XGBoost (d) K-Nearest-Neighbor (e) Logistic Regression, (f) Decision Tree Classifier, (g) AdaBoost, (h) Gradient Boost, (i) Gaussian Naive Bayes, **(j)* Proposed Stacked Ensemble**

REFERENCES

[1] Judith Balmaña, Laura Digiovanni, Pragna Gaddam, Michael F Walsh, Vijai Joseph, Zsofia K Stadler, Katherine L Nathanson, Judy E Garber, Fergus J Couch, Kenneth Offit, et al. 2016. Conflicting interpretation of genetic variants and cancer risk by commercial laboratories as assessed by the prospective registry of multiplex testing. *Journal of Clinical Oncology* 34, 34 (2016), 4071.

[2] Majed Dasouki, Omar Jawdat, Osama Almadhoun, Mamatha Pasnoor, April L McVey, Ahmad Abuzinadah, Laura Herbelin, Richard J Barohn, and Mazen M Dimachkie. 2014. Pompe disease: literature review and case series. *Neurologic clinics* 32 (2014), 751–776. Issue 3.

[3] William Gradishar, KariAnne Johnson, Krystal Brown, Erin Mundt, and Susan Manley. 2017. Clinical variant classification: a comparison of public databases and a commercial testing laboratory. *The oncologist* 22 (2017), 797–803. Issue 7.

[4] Steven M Harrison, Jill S Dolinsky, Amy E Knight Johnson, Tina Pesaran, Danielle R Azzariti, Sherri Bale, Elizabeth C Chao, Soma Das, Lisa Vincent, and Heidi L Rehm. 2017. Clinical laboratories collaborate to resolve differences in variant interpretations submitted to ClinVar. *Genetics in Medicine* 19 (2017), 1096–1104. Issue 10.

[5] Melissa J. Landrum, Jennifer M. Lee, Mark Benson, Garth R. Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Wonhee Jang, Karen Karapetyan, Kenneth Katz, Chunlei Liu, Zenith Maddipatla, Adriana Malheiro, Kurt McDaniel, Michael Ovetsky, George Riley, George Zhou, J. Bradley Holmes, Brandi L. Kattman, and Donna R. Maglott. 2018. ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Research* 46 (1 2018), D1062–D1067. Issue D1. https://doi.org/10.1093/nar/gkx1153

[6] Kirill Musin and Andrey Gaidel. 2020. Machine learning algorithms in the prediction of conflicts in clinical classification of genetic variants. *CEUR Workshop Proceedings*, 179–182.

[7] Giovanna Nicora, Susanna Zucca, Ivan Limongelli, Riccardo Bellazzi, and Paolo Magni. 2022. A machine learning approach based on ACMG/AMP guidelines for genomic variant classification and prioritization. *Scientific Reports* 12 (12 2022). Issue 1. https://doi.org/10.1038/s41598-022-06547-3

[8]]WinNT NIH. [n. d.]. What is a gene? https://medlineplus.gov/genetics/understanding/basics/gene/.

[9] Sue Richards, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, Wayne W Grody, Madhuri Hegde, Elaine Lyon, and Elaine Spector. 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 (2015), 405–423. Issue 5.

[10] Rosella Tomanin, Litsa Karageorgos, Alessandra Zanetti, Moeenaldeen Al-Sayed, Mitch Bailey, Nicole Miller, Hitoshi Sakuraba, and John J Hopwood. 2018. Mucopolysaccharidosis type VI (MPS VI) and molecular analysis: Review and classification of published variants in the ARSB gene. *Human mutation* 39 (2018), 1788–1802. Issue 12.

[11] V Venkata Durga Kiran, Sasumana Vinay Kumar, Suresh B Mudunuri, and Gopala Krishna Murthy Nookala. 2021. Comparative Study of Machine Learning Models to Classify Gene Variants of ClinVar. In *Data Management, Analytics and Innovation*. Springer, 435–443.

[12] Ingrid E C Verhaart and Annemieke Aartsma-Rus. 2019. Therapeutic developments for Duchenne muscular dystrophy. *Nature Reviews Neurology* 15 (2019), 373–386. Issue 7.

APPENDIX

<table>
<tr><td><b>1a</b></td><td>missense_Variant | synonymous_Variant | other_Variant</td></tr>
<tr><td><b>1b</b></td><td>unknown_Origin | germline_Origin | somatic_Origin | inherited_Origin | paternal_Origin | maternal_Origin | de-novo_Origin | biparental_Origin | uniparental_Origin | not-tested_Origin | tested-inconclusive_Origin | other_Origin</td></tr>
<tr><td><b>1c</b></td><td>has_Hereditary_cancer-predisposing_syndrome | has_Hereditary_breast_and_ovarian_cancer_syndrome | has_Familial_cancer_of_breast | has_Dilated_cardiomyopathy_1G |
has_Limb-girdle_muscular_dystrophy,_type_2J | has_Cardiovascular_phenotype |
has_Hypertrophic_cardiomyopathy | has_Ataxia-telangiectasia_syndrome |
has_Hereditary_nonpolyposis_colon_cancer | has_Dilated_Cardiomyopathy,_Dominant |
has_Familial_hypercholesterolemia | has_Breast-ovarian_cancer,_familial_2 |
has_Familial_adenomatous_polyposis_1 | has_Limb-Girdle_Muscular_Dystrophy,_Recessive |
has_Lynch_syndrome</td></tr>
<tr><td><b>1d</b></td><td>AA_REF | AA_ALT</td></tr>
<tr><td><b>1e</b></td><td>IE | IE_Loc</td></tr>
</table>