Machine Learning Models for Accurate Prioritization of Variants of Uncertain Significance

Daniel Mahecha[1], Haydemar Nuñez[2], Maria Lattig[1], and Jorge Duitama[2]

[1]SIGEN, Alianza Universidad de los Andes - Fundación Santa Fe de Bogota
[2]Universidad de los Andes

November 25, 2020

Abstract

The growing use of new generation sequencing technologies on genetic diagnosis has produced an exponential increase in the number of Variants of Uncertain Significance (VUS). In this manuscript we compare three machine learning methods to classify VUS as Pathogenic or No pathogenic, implementing a Random Forest (RF), a Support Vector Machine (SVM), and a Multilayer Perceptron (MLP). To train the models, we extracted 82,463 high quality variants from ClinVar, using 9 conservation scores, the loss of function tool and allele frequencies. For the RF and SVM models, hyperparameters were tuned using cross validation with a grid search. The three models were tested on a set of 5,537 variants that had been classified as VUS any time along the last three years but had been reclassified in august 2020. The three models yielded superior accuracy on this set compared to the benchmarked tools. The RF based model yielded the best performance across different variant types and was used to create VusPrize, an open source software tool for prioritization of variants of uncertain significance. We believe that our model can improve the process of genetic diagnosis on research and clinical settings.

Machine Learning Models for Accurate Prioritization of Variants of Uncertain Significance

Daniel Mahecha[1,2] Haydemar Nuñez[2], Maria Claudia Lattig[1,3], Jorge Duitama[2,*]

1. SIGEN, Alianza Universidad de los Andes - Fundación Santa Fe de Bogota,Colombia
2. Systems and Computing Engineering Department, Universidad de los Andes, Colombia
3. Facultad de Ciencias, Universidad de los Andes

* Corresponding author: ja.duitama@uniandes.edu.co

ABSTRACT

The growing use of new generation sequencing technologies on genetic diagnosis has produced an exponential increase in the number of Variants of Uncertain Significance (VUS). In this manuscript we compare three machine learning methods to classify VUS as *Pathogenic* or*No pathogenic* , implementing a Random Forest (RF), a Support Vector Machine (SVM), and a Multilayer Perceptron (MLP). To train the models, we extracted 82,463 high quality variants from ClinVar, using 9 conservation scores, the loss of function tool and allele frequencies. For the RF and SVM models, hyperparameters were tuned using cross validation with a grid search. The three models were tested on a set of 5,537 variants that had been classified as VUS any time along the last three years but had been reclassified in august 2020. The three models yielded superior accuracy on this set compared to the benchmarked tools. The RF based model yielded the best performance across different variant types and was used to create VusPrize, an open source software tool for prioritization of variants of uncertain significance. We believe that our model can improve the process of genetic diagnosis on research and clinical settings.

KEYWORDS

Variants of uncertain significance, variant interpretation, machine learning, pathogenicity prediction, genetic diagnosis

INTRODUCTION

Rare variants on genes involved in genetic disease produce a high toll of disability and premature death worldwide. For example, some variants on the *CFTR* gene cause cystic fibrosis (Strausbaugh & Davis, 2007), variants on the *HBB* gene can cause sickle cell disease (Kato et al., 2018), and variants on the *LDLR* gene cause familial hypercholesterolemia (Defesche et al., 2017), all diseases with heavy burdens on health as well as quality of life on patients and their families. Despite the low frequency of each genetic disorder, there are around eight thousand genes where single variants can lead to genetic diseases (Amberger, Bocchini, Scott, & Hamosh, 2020), resulting in a high total frequency of genetic diseases and affecting more than 300 million people worldwide. One of the main concerns on genetic diseases is diagnostic delay. For rare diseases, 80% of which have a known genetic cause, the delay until a correct diagnosis is given is on average 4.8 years (Evans, 2018) but can be as long as 30 years (Gainotti et al., 2018), causing an additional burden of stress for medical practitioners, patients, and their families.

The process of genetic diagnosis aims to correctly identify the genetic variant that is causing a specific disease. This is a complex process that involves taking into account multiple data sources including, but not limited to, gene and phenotype association, allele frequencies on a population relevant to the patient, the inheritance pattern of the disease, functional studies related to suspected variants, and computational predictions (Richards, et al., 2015). Until recently, the process was based on gene panels or chromosomal arrays that included a limited number of variants known to be pathogenic and associated with particular diseases (Fogel, Satya-Murti, & Cohen, 2016; Miller, et al., 2010). With this approach, variants that are not included in the assay cannot be detected. In the last ten years, the introduction of high throughput sequencing technologies (Whole Exome and Whole Genome Sequencing) have increased the yield of variants detected in a single test, and have demonstrated superior clinical and diagnostic utility than the formerly used first line-tests for many diseases (Clark, et al., 2018). However, due to the complexity of the process, the higher yield of detected variants has not been coupled with a proportional increase in variant interpretation capabilities, resulting in an explosion of variants of uncertain significance (VUS). In fact, the number of variants classified as VUS have exponentially increased in the last few years, and the majority of clinically-interpreted variants are currently VUS (Weile & Roth, 2018). This problem is even more prevalent among "underrepresented minorities" compared to Caucasian populations, as there are fewer genomic and clinical studies with patients on these populations (Walsh, et al., 2019).

Ideally, VUS are reclassified in a more informative category (*pathogenic* , *likely pathogenic* , *likely benign* , or*benign*) but achieving this goal requires ascertaining new information on the variant through experimental or population studies, which take time and consume resources. One way to prioritize VUS with a higher probability of being pathogenic (i.e. disease-causing) for further studies is to use computational predictive tools. Computational predictive tools are models that estimate the probability that a given variant is deleterious or pathogenic based on information about its evolutionary conservation, its effect on protein structure or function (if it is a coding variant), or its effect on relevant features of the DNA sequence (v.g. splice sites, regulatory sites, protein-DNA binding sites, among others). The most commonly used tools (Ghosh, Oak, & Plon, 2017) yield scores ranging from 0 to 1, some of which reflect a probability value, while not all are calibrated to reflect a true probability. Some tools such as CADD (Rentzsch,, et al., 2019) yield phred-scores as well. Probability scores can be obtained using calibration formulas. Using these probability scores researchers and clinicians are able to prioritize VUS with a higher probability of being pathogenic (i.e. disease causing), and can potentially guide clinical decision-making processes for these types of variants when additional evidence is lacking. However, currently used predictors have several shortcomings. First, most predictors are designed for missense type variants, leaving out an important proportion of the variants currently classified as VUS, which have different consequence data types (Figure 1). Some frameworks that work for other variant types (v.g. MutationSVM), have separate tools for each variant type. Additionally, tools for missense variants tend to overestimate the pathogenicity of benign variants. Finally, while other

tools perform better as classifiers of pathogenic vs. non pathogenic, the probability distributions for VUS do not reflect the suggested thresholds of probability suggested by the ACMG for variant classification on the four remaining categories.

Here, we present a comparison of three machine learning models (Random Forest, Support Vector Machine, and a Five-Layer Perceptron) in a one-for-all approach, meaning that each model can correctly prioritize variants of different consequence types. To increase their predictive power and interpretability, we merged ACMG *Benign* and*Likely Benign* categories into a unique Benign category, and ACMG*Pathogenic* and *Likely Pathogenic* categories into a unique Pathogenic category. To avoid circularity bias, we trained our models using conservation scores that did not include clinical interpretation data. Additionally, we demonstrated that including allele frequencies increases the predictive power of the models. To assess the performance of the resulting models for prioritization of the VUS population, we benchmarked the resulting models against currently used predictors using a set of variants that had been classified as VUS on the last three years, but have been reclassified into the remaining categories as of august 2020 on ClinVar (Landrum, et al., 2017), showing superior performance among different variant consequence types.

RESULTS

In order to achieve the goal to prioritize VUS with a higher probability of being pathogenic and overcome the limitations of current predictor tools, we developed three models comparing three machine learning approaches (Random Forest, Support Vector Machine and a Neural Network with a Five-layer multilayer perceptron architecture). Models were trained with a set of 82,426 high quality variants from the ClinVar database and tested with a set of variants that had been classified as VUS anytime during the last three years, but had been reclassified with high confidence in any of the 4 informative categories (Pathogenic, Likely pathogenic, Likely Benign, Benign). To increase the size of the training set and ease the interpretation of results we merged the Pathogenic and Likely pathogenic categories into a unique Pathogenic category, and the Benign and Likely Benign category into a unique Benign category.

Machine learning models to improve classification of VUS

We explored three different machine learning strategies to classify variants that are currently assigned as variants of uncertain significance (VUS) by standard variant interpretation pipelines.

After building three models for VUS pathogenicity prediction based on a Random Forest (RF), a Support Vector Machine (SVM), and a Five-Layer Perceptron (MLP), their performance was measured on a set of variants previously classified as VUS but reclassified in any of the other categories in ClinVar with at least two quality stars. This set includes 5,537 variants representative of the main variant consequence types (Figure 1a), including 2,008 (36.3%) *missense* variants, 1,844 (33.3%) *synonymous* variants, 349 (6.3%) *intron* variants, 475 (8.6%) *splice* variants, 340 (6.32%) non-coding mRNA variants, 69 (1.25%) coding INDEL variants, 151 (2.73%) intergenic, and 290 (5.22%) of other variant types (5-prime UTR variants, 3-prime UTR variants, upstream gene variants, downstream gene variants, TF binding site variants, and nonsense variants). As measured by the area under the curve of the Receiving Operator Characteristic curves (AUROC), our three models outperform the best performing of the benchmarked tools (CADD, with an AUC of 0.92), with an AUC of 0.97 for the RF and the MLP based models, and a AUC of 0.96 for the SVM based model (Figure 1b). Additionally, the three models were trained separately including and excluding 1000 Human Genomes global allele frequencies to compare their performance on the original test set. For all the models analyzed, including the 1KG Global Allele Frequencies increased performance measured by the AUC (See Supplementary Figure S1).

High performance among different variant consequence types

Most of the currently available tools, v.g. SIFT (Vaser, Adusumalli, Leng, Sikic, & Ng, 2016), PolyPhen (Adzhuvei, et al., 2010), and Revel (Ioannidis, et al., 2016) are designed to yield scores for missense type variants exclusively, resulting in a lower performance in the dataset as it includes diverse variant consequence types. Only a third of the variants of this dataset are missense type, and there are significant numbers of

synonymous, non coding transcript exon, intron, and splice variants. There is a smaller number of other consequence types as well, namely frameshift and nonsense variants. For this reason, to get a fairer representation of the performance of the trained models against the benchmarked tools, the ROCs for the same models were plotted for the subsets of the specific variant consequence types. On the subset of missense variants (Supplementary Figure S2a), the AUC of the RF and the MLP (0.97) outperform the SVM (0.96). As most tools are designed for this consequence type, there is a general superiority of the AUCs compared with the other variant types. Revel yields an AUC of 0.96, equal to the SVM and slightly lower than the MLP and RF. The commonly used SIFT and PolyPhen had lower AUCs than other the analyzed tool (0.81 and 0.85, respectively). M-CAP (AUROC=0.95) , MetaLR, and MetaSVM (AUROC=0.93) yield high accuracy on missense variants as well.

For the splice type variants (Supplementary Figure S2b), our RF yields an AUC of 0.97, the MLP an AUC of 0.93, and the SVM an AUC of 0.90. The CADD tool yields an AUC of 0.95 outperforming our MLP and SVM. For synonymous variants (Supplementary Figure S2c), the AUROCs are consistently lower. In these variants, our three models get an AUC of 0.89, outperforming CADD (AUROC=0.57). For non-coding mRNA variants (Supplementary Figure S2d), the AUROCs of our models are outperformed by CADD. The RF yielded an AUROC 0f 0.89, the SVM of 0.85, and the MLP, of 0.89, lower than CADD with an AUROC of 0.93. For the intron type variants (Supplementary Figure S2e), the RF yields an AUROC of 0.89, the SVM of 0.84, and the MLP of 0.83. The CADD score yielded an AUROCC of 0.76. Our models misclassify coding INDEL variants (Supplementary Figure S2f), showing AUROCs lower than 0.5, while CADD has an AUROC of 0.78. In the case of intergenic variants (Supplementary Figure S2g), the three models yielded an AUROC of 0.58, while CADD yielded an AUROC of 0.89. For other variant types (Supplementary Figure S2h), performance is better, with AUROC=0.92 for the RF and the SVM, and AUROC=0.89 for the MLP. In this variant type, CADD AUROC=0.95.

Improved performance through ensembling with CADD

To overcome the shortcomings of the developed models on certain variant types (namely *non-coding mRNA* variants, *coding INDEL*variants, *intergenic* variants, and other types), and considering that ensemble approaches have shown increased performance (Ghosh, Oak, & Plon, 2017), we retrained the models using the CADD score as an additional feature. Compared to other variant deleteriousness prediction tools, CADD does not use clinical variants for its training, so using it avoids the circularity bias that would arise from using other tools like REVEL or PolyPhen in our training and testing with a ClinVar variant population. The models were trained in the same fashion than described above, tuning the hyperparameters with cross validation using a grid search approach. The overall performance of the models improved, as seen on Figure 1d. The RF and MLP based models yielded an AUROC of 0.98, and the SVM model of 0.97. As seen on Figure 1c, the highest improvement is seen on coding INDELs and intergenic variants, and a more modest increase in the AUROC for splice, non coding mRNA, and other variant types. Synonymous variants experienced a decrease in accuracy as measured by the AUROCs. A profiling analysis showed that virtually all synonymous variants are labelled as Benign, while virtually all Frameshift variants are labelled as Pathogenic, implying that the models assign the benign label to all synonymous, and the pathogenic label to all frameshift variants. For the variant consequence types analyzed, synonymous variants yield the lowest results on AUCs. Moreover, the current classification of non-VUS synonymous variants on ClinVar (99% are classified as Bening) is not matched by the CADD scores which predict a much higher number to be pathogenic (Supplementary Figure S2).

As shown in figure 2a, on the subset of missense variants, the AUC of the RF (0.97) outperforms the SVM (0.96) and the MLP (0.96), and all the benchmarked tools. For the splice type variants (Figure 2b), our RF yielded an AUC of 0.99; the SVM, of 0.97; and the MLP, of 0.98. CADD yielded an AUC of 0.95; our MLP an AUC of 0.96; and the ada score and AUC of 0.95. For synonymous variants (Figure 2c), the RF yielded an AUC of 0.79. Non coding exon type variants (Figure 2d), the AUCs are slightly lower. The RF yielded an AUC of 0.98, the SVM and the MLP, of 0.96 and 0.97, respectively, higher than CADD with 0.93. For the intron type variants (Figure 3e), the RF yielded an area of 0.92. In all cases except coding INDEL

variants (Figure 2f), our models yield AUCs higher than CADD, and the RF based model gets the highest performance across the assessed variant types.

Score distributions for currently used tools

The distribution of values of the retrieved features, many of which are currently used as deleteriousness/pathogenicity prediction scores, were plotted for Benign and Pathogenic variants, as well as for Variants of Uncertain Significance. Figure 3 shows the distribution of values for three of the most commonly used tools, namely SIFT, PolyPhen and Revel. Considering that the SIFT score assigns a 0 value to deleterious variants, in contrast with the typical score value of 1 for deleterious/pathogenic variants, its histogram was plotted using the 1-SIFT value to allow for easier comparison with the other tools. As seen on Figure 3, 1-SIFT scores have a great proportion of values [?] 1 for Benign variants, suggesting an overestimation of deleteriousness. PolyPhen scores have values [?] 1 for benign, and values [?] 0 for pathogenic variants as well. However, for VUS variants SIFT and PolyPhen have pronounced distributions with peaks on the extreme values, while the Revel scores have a less markedly bimodal distribution. An ideal prediction score for VUS variants would classify them on two clear clusters (in a similar way to PolyPhen) while avoiding classification errors.

Distribution of the developed models on current VUS

Finally, we ran our models on a set of ClinVar variants currently classified as VUS (Figure 4). The Random Forest model predicts that approximately three quarters of the variants to be pathogenic and one quarter benign. On the distribution of probability of pathogenicity, there is a sharp peak in values of probability of pathogenicity close to 1, and a less marked peak in values around thirty percent. The support vector machine model predicts that approximately two thirds of the variants to be pathogenic and one third benign. On the distribution of probability of pathogenicity, there is a sharp peak in values of probability of pathogenicity close to 1, and a less marked peak in values close to zero. The multilayer perceptron model predicts approximately three quarters of the variants to be pathogenic and one quarter benign. Its distribution of probability behaves in a similar fashion to the SVM model.

DISCUSSION

The interpretation of genomic variation data in clinical settings has been one of the biggest challenges in achieving a successful use of next generation sequencing in medical practice. Given the relevance of this process for current diagnosis of genetic diseases, variant interpretation is one of the most important topics in current bioinformatics. To develop this model, we combined different machine learning techniques and scores from widely used tools, as an effort to try to improve the accuracy of classification of Variants with Uncertain Significance (VUS). Our models showed improved accuracy compared to current solutions analyzing data from a large set of variants previously classified as VUS and distributed over different consequence types..

VUS raises concerns for both patients and for clinicians working on genetic diagnosis. From the patient perspective, genetic testing can yield or confirm a diagnosis, inform the probability of developing a disease for the patient or relatives, and, if the variant is actionable, offer a possibility of treatment. Thus, the uncertainty created by VUS on genetic testing can lead to a variety of emotional responses in patients. Some of the most reported answers to a VUS result are stress and distress, both on patients and on their relatives. Additionally, VUS results are more difficult to understand as patients tend to have a more deterministic view of genetics, and in many scenarios tend to misinterpret VUS results as more similar to a negative result (Clift, et al., 2019). Moreover, due to the variability of patient medical and psychosocial contexts, there is no consensus on clinical best practices to handle VUS results. Some propose to withhold them from patients, so clinicians and labs have to deal with them on a case-by-case basis focusing on pre and post diagnosis counseling to minimize potential harm (Hoffman-Andrews, 2017).

From the clinician point of view, a VUS result raises concerns on how to counsel the affected patient and their family, and how it might change the clinical management. The ACMG guidelines state that a VUS should not be used as part of clinical decision making. Therefore, it is advised that, whenever feasible, the clinician

should pursue additional efforts to classify the variant. Additional monitoring and tracing of the patient might be needed if the variant is reclassified (Hoffman-Andrews, 2017). In any case, these efforts involve important time and monetary investments: they can be directed at the patient and family level (i.e. testing for the variant on parents and other relatives) or, if the laboratory has research facilities, functional studies to validate the variant consequence. Other approaches include the work by Sun, et al. (2020), which aims to tackle this problem by proactively creating comprehensive maps of cell-based assays for the missense variants of specific genes, or the work by Walsh, et al. (2019), which compares variant frequency between patient cohorts and reference population cohorts. However, so far these approaches are available for a number of selected genes and diseases. Thus, in resource limited settings VUS prioritization is a paramount need and our tools can help us select VUS with the highest probability of being pathogenic with a high accuracy.

As demonstrated by Liu, Wu, Li, & Boerwinle, et al. (2016), combining information of several predictive scores increases the predictive accuracy of missense variant classification. Here, we show that combining the information of high accuracy conservation-based variant deleteriousness tools like CADD, SIFT, and Eigen (Ionita-Laza, McCallum, Xu, & Buxbaum, 2016) yields improved accuracy across a variety of variant types including *missense* , *splice* , *intron* ,*intergenic* , and *synonymous* . However, synonymous variants obtained the less accurate results with both our models and CADD (Figure 3). Recent work suggests that synonymous variants might be more deleterious than would be predicted from current clinical significance annotations (Zeng & Bromberg, 2020). We believe that additional research might be needed to ascertain the true pathogenic potential of these class of variants.

METHODS

Selection of variants for model training and testing

From ClinVar, we selected 82,463 variants which had at least two quality stars (i.e. assertion criteria available, multiple submitters, and no conflicts in the interpretation) and were not classified as VUS in the ClinVar database, version 08/03/2020. We sampled the ClinVar database versions from 06/15/2017, 12/03/2017, 06/03/2018, 12/02/2018, 06/03/2019, 12/06/2019, to look for variants that were classified as VUS on those dates, but had been reclassified on any of the four remaining categories (*pathogenic* , *likely pathogenic* , *likely benign* , or *benign*) and were included on the group of 82,463 variants, finding 5,537 variants that were reserved for further benchmarking as the *ex-VUS* set. To increase predictive power by including more variants for training, and ease the interpretability of the results, *Bening* and *Likely Benign* variants were merged into a unique *Benign* label, and *Pathogenic* and*Likely Pathogenic* variants were merged into a unique*Pathogenic* label.

Variant feature selection

To assess possible attributes for model training, we used 24 variant features, including splice site predictors, conservation scores, deleteriousness/pathogenicity scores, allele frequency, and consequence type, from the Ensembl Variant Effect Predictor (McLaren, et al., 2016; Zerbino, et al., 2018)). Features with high Pearson correlation were depurated. Additionally, features with values coming from models trained with clinical significance data were discarded to avoid circularity biases on our model estimation phase. First, the features used for training were: ada score, codon degeneracy score, integrated fitness conservation score, BLOSUM62 score, Eigen score, phyloP score, Gerp score, SIFT score, the Loss of Function tool score, the allele frequencies from the 1000 human genomes project global dataset, and the variant consequence type codified as dummy binary variables. Clinical Significance was used as the label for training, and codified using*1* for *pathogenic,* and *0* for *benign* . To correct for class unbalance (2/3 benign vs. 1/3 pathogenic variants) we randomly undersampled benign variants to equalize the number of*pathogenic* variants. After testing for the models performance on the ex-VUS set, models were retrained with the procedure described before, adding the CADD phred score (retrieved from Ensembl Variant Effect Predictor) as a feature for the variants.

Parameter tuning for different machine learning models

The dataset of 76,926 variants was split on a training and a test set with a 80:20 ratio. Using the SciKit Learn library for Python 3, we trained a model based on a Random Forest, and a model based on a Support Vector Machine with a RBF kernel. Hyperparameters were tuned using a grid search with cross-validation approach optimizing the area under the ROC curves. The hyperparameters tuned on the Random Forest were:*Maximum depth, selection criteria, and number of estimators* . The hyperparameters tuned on the Support Vector Machine were *C value and gamma value* . Additionally, a Five-Layer Perceptron was trained using a batch size of 50 and 25 epochs on the Keras library for Python 3 with a TensorFlow backend. A ReLu was chosen as an activation function for the hidden layers and a Sigmoid as a function for the output layer. To assess model performance, we plotted the area under the ROC curves and calculated the area under the ROC curves. We compared the models trained including and excluding the 1000 human genomes project global allele frequencies.

Finally, we further tested the resulting models on the set of 5,537*ex-VUS* and compared their performance against the scores of commonly used prediction tools (retrieved from Ensemble VEP). First, we tested the model on the whole set of variants irrespective of their consequence type. Then, to make a fairer assessment against tools that yield scores only for specific consequence types (such as*missense* type variants), we plotted the ROC curves and calculated AUCs the same models but on the specific subpopulations of variants.

SOFTWARE AVAILABILITY STATEMENT

Software and source code available from: https://github.com/danielhmahecha/VusPrize

DATA AVAILABILITY STATEMENT

All data analyzed in this study is publicly available at the ClinVar database of NCBI (*https://www.ncbi.nlm.nih.gov/clinvar/*). Scores related to each variant were downloaded from the Variant Effect Predictor web service of Ensembl (https://www.ensembl.org/info/docs/tools/vep/index.html).

CONFLICT OF INTEREST STATEMENT

All authors declare that they do not have any conflict of interest related to the research presented in this manuscript.

ACKNOWLEDGEMENTS

Internal funding from Universidad de los Andes through the SIGEN agreement led by MCL and the FAPA initiative led by the Vice-presidency of Research and Knowledge Creation and awarded to JD. The authors would like to thank Fabian Heredia and Yacir Ramirez for discussions at the early stages of this project.

REFERENCES

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., ... Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature methods*,*7*(4), 248–249. https://doi.org/10.1038/nmeth0410-248 Amberger, J. S., Bocchini, C. A., Scott, A. F., & Hamosh, A. (2019). OMIM.org: leveraging knowledge across phenotype-gene relationships.*Nucleic acids research*, *47*(D1), D1038–D1043. https://doi.org/10.1093/nar/gky1151 Clark, M. M., Stark, Z., Farnaes, L., Tan, T. Y., White, S. M., Dimmock, D., & Kingsmore, S. F. (2018). Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases. *Npj Genomic Medicine*, *3*(1), 1–10. https://doi.org/10.1038/s41525-018-0053-8 Clift, K., Macklin, S., Halverson, C., McCormick, J. B., Abu Dabrh, A. M., & Hines, S. (2020). Patients' views on variants of uncertain significance across indications. *Journal of Community Genetics*,*11*(2), 139–145. https://doi.org/10.1007/s12687-019-00434-7 Defesche, J. C., Gidding, S. S., Harada-Shiba, M., Hegele, R. A., Santos, R. D., & Wierzbicki, A. S. (2017). Familial hypercholesterolaemia. *Nature Reviews. Disease Primers*, *3*, 17093. https://doi.org/10.1038/nrdp.2017.93 Evans W. R. (2018). Dare to think rare: diagnostic delay and rare diseases. *The British journal of general practice : the journal of the Royal College of General Practitioners*, *68*(670), 224–225. https://doi.org/10.3399/bjgp18X695957 Fogel, B. L., Satya-Murti, S., & Cohen, B. H. (2016). Clinical exome sequencing in neurologic disease. *Neurology. Clinical practice*,*6*(2), 164–176.

https://doi.org/10.1212/CPJ.0000000000000239 Gainotti, S., Mascalzoni, D., Bros-Facer, V., Petrini, C., Floridia, G., Roos, M., ... Taruscio, D. (2018). Meeting Patients' Right to the Correct Diagnosis: Ongoing International Initiatives on Undiagnosed Rare Diseases and Ethical and Social Issues. *International journal of environmental research and public health*, *15*(10), 2072. https://doi.org/10.3390/ijerph15102072 Ghosh, R., Oak, N., & Plon, S. E. (2017). Evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines. *Genome Biology*, *18*(1), 1–12. https://doi.org/10.1186/s13059-017-1353-5 Hoffman-Andrews, L. (2017). The known unknown: The challenges of genetic variants of uncertain significance in clinical practice. *Journal of Law and the Biosciences*, *4*(3), 648–657. https://doi.org/10.1093/jlb/lsx038 Ioannidis, N. M., Rothstein, J. H., Pejaver, V., Middha, S., McDonnell, S. K., Baheti, S., ... Sieh, W. (2016). REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *American Journal of Human Genetics*, *99*(4), 877–885. https://doi.org/10.1016/j.ajhg.2016.08.016 Ionita-Laza, I., Mccallum, K., Xu, B., & Buxbaum, J. D. (2016). A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nature Genetics*, *48*(2), 214–220. https://doi.org/10.1038/ng.3477 Kato, G. J., Piel, F. B., Reid, C. D., Gaston, M. H., Ohene-Frempong, K., Krishnamurti, L., ... Vichinsky, E. P. (2018). Sickle cell disease. *Nature reviews. Disease primers*, *4*, 18010. https://doi.org/10.1038/nrdp.2018.10 Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., ... Maglott, D. R. (2018). ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, *46*(D1), D1062–D1067. https://doi.org/10.1093/nar/gkx1153 Liu, X., Wu, C., Li, C., & Boerwinkle, E. (2016). dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Human Mutation*, *37*(3), 235–241. https://doi.org/10.1002/humu.22932 McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., ... Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, *17*(1), 1–14. https://doi.org/10.1186/s13059-016-0974-4 Miller, D. T., Adam, M. P., Aradhya, S., Biesecker, L. G., Brothman, A. R., Carter, N. P., ... Ledbetter, D. H. (2010). Consensus Statement: Chromosomal Microarray Is a First-Tier Clinical Diagnostic Test for Individuals with Developmental Disabilities or Congenital Anomalies. *American Journal of Human Genetics*, *86*(5), 749–764. https://doi.org/10.1016/j.ajhg.2010.04.006 Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., & Kircher, M. (2019). CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*, *47*(D1), D886–D894. https://doi.org/10.1093/nar/gky1016 Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., ... Rehm, H. L. (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 Strausbaugh, S. D., & Davis, P. B. (2007). Cystic Fibrosis: A Review of Epidemiology and Pathobiology. *Clinics in Chest Medicine*,*28*(2), 279–288. https://doi.org/10.1016/j.ccm.2007.02.011 Sun, S., Weile, J., Verby, M., Wu, Y., Wang, Y., Cote, A. G., ... Roth, F. P. (2020). A proactive genotype-to-patient-phenotype map for cystathionine beta-synthase. *Genome medicine*, *12*(1), 13. https://doi.org/10.1186/s13073-020-0711-1 Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M., & Ng, P. C. (2016). SIFT missense predictions for genomes. *Nature Protocols*,*11*(1), 1–9. https://doi.org/10.1038/nprot.2015.123 Walsh, R., Mazzarotto, F., Whiffin, N., Buchan, R., Midwinter, W., Wilk, A., ... Ware, J. S. (2019). Quantitative approaches to variant classification increase the yield and precision of genetic testing in Mendelian diseases: The case of hypertrophic cardiomyopathy. *Genome Medicine*, *11*(1), 1–18. https://doi.org/10.1186/s13073-019-0616-z Weile, J., & Roth, F. P. (2018). Multiplexed assays of variant effects contribute to a growing genotype–phenotype atlas. *Human Genetics*, *137*(9), 665–678. https://doi.org/10.1007/s00439-018-1916-x Zeng, Z., & Bromberg, Y. (2019). Predicting Functional Effects of Synonymous Variants: A Systematic Review and Perspectives. *Frontiers in genetics*, *10*, 914. https://doi.org/10.3389/fgene.2019.00914 Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., ... Flicek, P. (2018). Ensembl 2018. *Nucleic Acids Research*, *46*(D1), D754–D761. https://doi.org/10.1093/nar/gkx1098

FIGURE LEGENDS

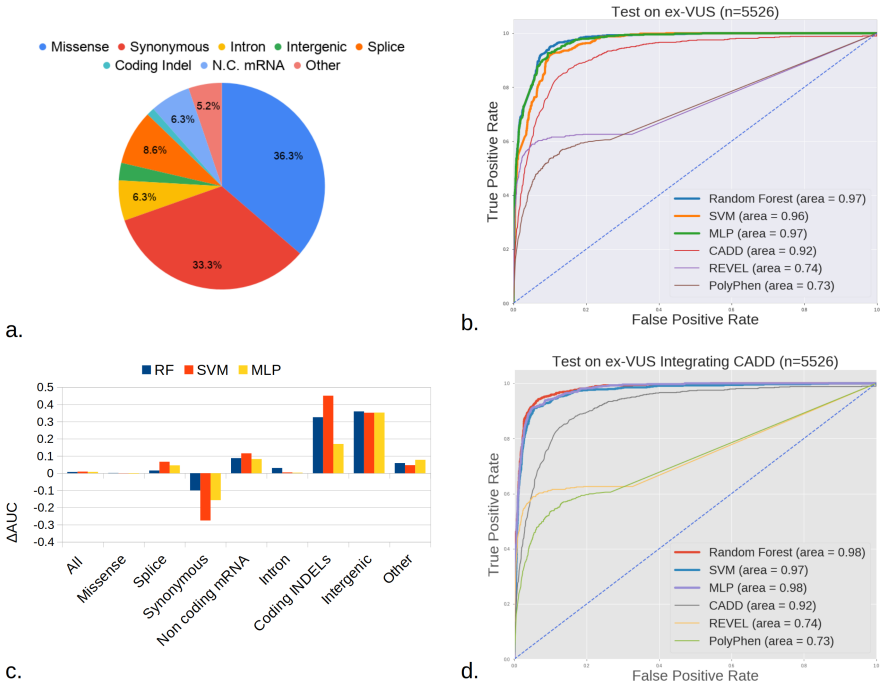
Figure 1. a) Pie chart showing the composition of the ex-VUS set by variant consequence type. b) Receiver Operating Characteristic (ROC) curves for our initial Random Forest, Support Vector Machine (SVM), and

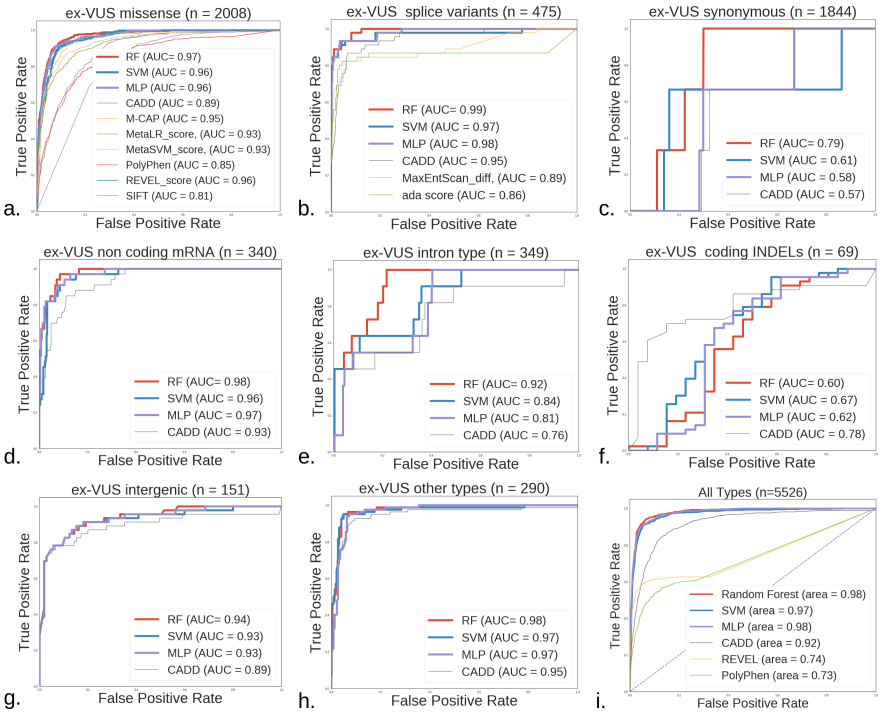
Multilayer Perceptron (MLP) based models, and CADD, REVEL, and PolyPhen. Our models are drawn with thicker lines. c) Bar graph showing the change in the Area Under the Curve (ΔAUC) for each variant consequence type in the ex-VUS sample, comparing the models developed before and after including CADD as a feature for model training. d) Receiver Operating Characteristic (ROC) curves for Random Forest, Support Vector Machine (SVM), and Multilayer Perceptron (MLP) based models trained including CADD phred score as a feature, and CADD, REVEL, and PolyPhen. Our models are shown with thicker lines.

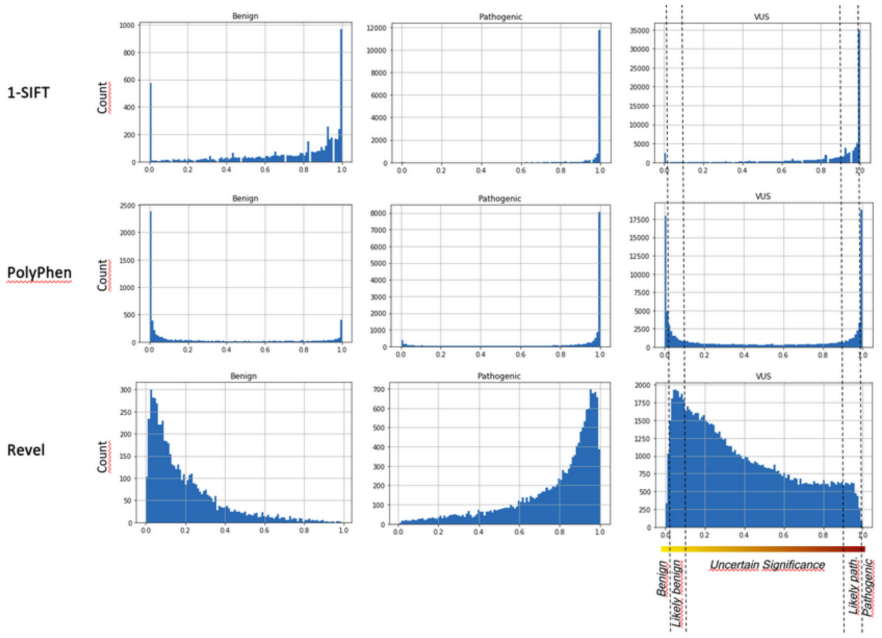
Figure 2. Receiver Operating Characteristic (ROC) curves for the Random Forest (RF), Support Vector Machine (SVM), and Multilayer Perceptron (MLP) models trained including CADD phred score as a feature. Curves for our models are shown in thicker lines along with benchmarked scores for a) *missense* , b)*splice* , c) *synonymous* , d)*non-coding mRNA* , e) *intron* , f), *coding INDEL* , g)*intergenic* , h) other variant types, and i) all variant consequence types.

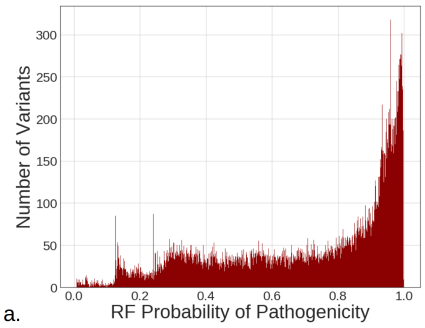
Figure 3. Distribution of values of 1-SIFT, PolyPhen, and Revel scores for Benign AND Pathogenic variants, and variants of Uncertain Significance. For the Variants of Uncertain Significance, the thresholds for each of the ACMG categories are displayed.

Figure 4. Distribution of probability of pathogenicity values for variants currently classified as Variants of Uncertain Significance on ClinVar for a) the Random Forest (RF) based model, b) the Support Vector Machine (SVM) based model, and c) the Multilayer Perceptron (MLP) based model. Additionally, d) shows a pie chart of the predictions of the RF based model for the current VUS on ClinVar.

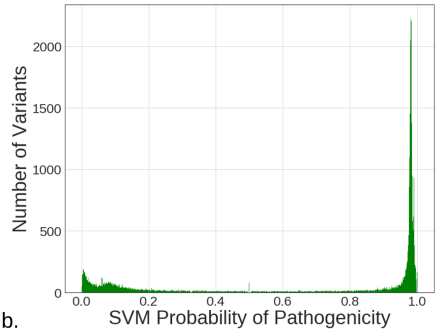




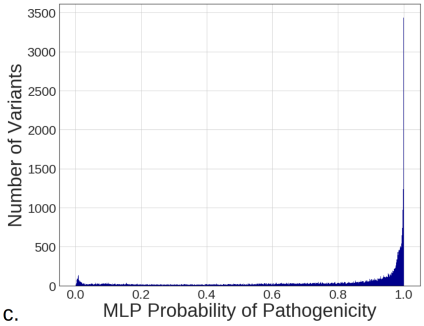




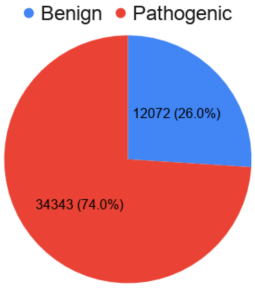
a.



b.



c.



d.