 doi:10.20944/preprints202011.0519.v1

Article

AmazonForest: In-silico meta-prediction of pathogenic variants

Helber Palheta [1], Wanderson Gonçalves e Gonçalves [1,2], Leonardo Miranda de Brito [1], Arthur Ribeiro dos Santos [1], Marlon dos Reis Matsumoto [1], Ândrea Ribeiro-dos-Santos [1,2] and Gilderlanio Santana Araújo [1,*]

[1] Laboratory of Human and Medical Genetics, Institute of Biological Sciences, Graduate Program of Genetics and Molecular Biology, Federal University of Pará, Belém 66075-110, Brazil; gilderlanio@gmail.com

[2] Research Center on Oncology, Graduate Program of Oncology and Medical Science, Federal University of Pará, Belém 66073-000, Brazil; akelyufpa@gmail.com

* Correspondence: gilderlanio@gmail.com

Abstract: ClinVar is a web platform that stores around 774k curated entries, which allows exploring genetic variants and their associations with complex phenotypes. A partial set of ClinVar's genetic associations were reported with conflict of interpretation or uncertain clinical impact significance, which currently challenges clinicians and geneticists. Here, we evaluate the performance of data pre-processing methods combined with classical prediction methods, such as Naive Bayes, Random Forest, and Support Vector Machine to build a meta-prediction model aiming to improve genetic pathogenicity interpretation. Models were trained with ClinVar data (September 2020), and genetic variants were annotated with eight functional impact predictors catalogued with SnpEff/SnpSift (v4.3). A 10-fold cross-validation strategy was performed for evaluation by accuracy, F1-Score, Receiver Operating Characteristic, Area Under Curve. The best meta-prediction model raises by combining one-hot encoding with tree-based classifiers as Random Forest, which shows Area Under Curve $\geq$ 0,93. We predict pathogenicity for 109k genetic variants, which were found labeled as uncertain significance or conflict of interpretation. Additionally, we implemented AmazonForest (https://www.lghm.ufpa.br/amazonforest), a web tool to query data for a set of 5k variants that were predicted with high pathogenic probability (RF_{prob} >= 0.9).

Keywords: Meta-prediction, Encoding data, ClinVar, Classification, Random Forest, Naive Bayes, Support Vector Machine

1. Introduction

Next-generation sequencing (NGS) methods have allowed genome-wide analyses of the human genome. Genome-wide association studies (GWAS) and candidate gene studies produced a large volume of genetic associations between single nucleotide polymorphisms (SNPs), insertions/deletions (INDELs) and complex diseases. Most of these associations show variable effects and genetic diversity among populations [1,2]. Variants with highly pathogenic effects have been shown to be responsible for the development of several types of cancer [3], Type 2 diabetes [4], and Alzheimer's disease [5,6]. Understanding the biological role and impact of these variants at a clinical and personalized level is a complex task.

ClinVar is an online database that stores around 774,000 curated entries that show relationships between phenotypes and genetic variants (SNPs or INDELs) and their clinical relevance (classified as either benign or pathogenic) [7]. ClinVar has improved our understanding of the functional role of genetic variants as research increasingly focuses towards precision medicine [8]. However, many genetic variants are yet to be functionally classified and remain with conflicting interpretations (CI) or with uncertain significance (VUS).

Distinct machine learning (ML) meta-prediction models have been proposed for pathogenicity prediction of new genetic variants. Generally, each meta-prediction models have been suggested

 doi:10.20944/preprints202011.0519.v1

for the analysis of a single variant class (synonymous or non-synonymous variants) [9–13] and most meta-predictors were used for the pathogenicity prediction of VUS and CI variants [9,10,12,13]. Interestingly,the majority of the recently proposed meta-predictors are based on decision trees or ensembles of decision trees, which constitute models with clear interpretations. Ensemble-based methods, such as Random Forest, are promising for pathogenicity prediction of coding and non-coding variants [9–11,13]. However, these models have shown differences regarding data training methods, specifically on the number of features used to build each classification models.

To the best of our knowledge, the performance of data encoding methods combined with machine learning techniques has not yet been assessed when training with categorical data or handling missing data. Thus, this study focused on building a meta-predictor model based on categorical data for pathogenicity interpretation. Here, we proposes a pathogenicity meta-predictor based on a dataset of 42,000 records without missing data. Genetic variants were functionally annotated by eight functional predictors that show only categorical data. We investigated the performance of encoding strategies for categorical data with classical machine learning methods, such as Naive Bayes, Random Forest, and Support Vector Machine. Among 167,412 identified variants with data for all predictors, only 42,813 were classified as benign or pathogenic, and the remaining were classified as uncertain significance (VUS) or with interpretation conflict (CI). Our best model was applied to reclassify 124,000 genetic variants.

2. Materials and Methods

2.1. ClinVar data

ClinVar data in stored in a .vcf file with a high volume of genetic variants of clinical importance (obtained on October 2, 2020. https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh38/). Initially, the dataset showed 774,921 genetic variants. Each variant is classified according to the ACMG-AMP[14] which labels, wich correspond to the following categories: benign, likely benign, variant of uncertain significance, likely pathogenic, pathogenic or interpretation conflict. For our data pre-processing, we grouped these classes into four labels: i) benign/likely benign into benign; ii) pathogenic/likely pathogenic into pathogenic; iii) variant of uncertain significance (VUS); iv) Conflicting interpretations (CI). Subsequently, the variant dataset was dived into: Benign (N = 263,742), Pathogenic (N = 124,299), VUS (N = 345,246), and CI (N = 41,634).

2.2. Functional impact annotation

SnpEff and SnpSift (v.4.3) configured with dbNSFP4.0 were used to extract prediction data from ClinVar variants. Hence, our meta-predictor was built based on categorical data extracted from seven *in silico* predictors: FATHMM, SIFT, Polyphen-2, PROVEAN, MutationAssessor, MutationTaster2, and LRT. Each predictor is independent and based on distinct approaches such as sequence characteristics, conservation, and amino acid changes. All predictors are described in detail as follows:

- FATHMM predicts the functional effects of coding and non-coding variants. It is done through the combination of wild type and mutated sequences in a hidden Markov model, which identifies mutations in peptide chains, showing the alignment of homologous sequences and conserved protein domains [15].
- SIFT (Sorting Intolerant From Tolerant) is a prediction tool that codes an algorithm for amino acid substitution analysis. It assumes that important positions in a protein sequence have been conserved throughout evolution and therefore substitutions at these positions may affect protein function. The algorithm sorts changes in a polypeptide chain as tolerant or intolerant according to its evolutionary conservation [16].
- Polyphen-2 (Polymorphism Phenotyping v2) predicts the impact of amino acid substitutions on structural stability, physical interactions, and human proteins function. The probability of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 November 2020 doi:10.20944/preprints202011.0519.v1

a mutation being pathogenic is based on the extraction of sequence annotations, structural attributes, and conservation profiles in protein-coding regions [17].

- PROVEAN (Protein Variation Effect Analyzer) is a predictor that provides a generalized approach to predict the functional effects on variations in a peptide chain. These effects include SNPs, INDELs, or multiple amino acid substitutions. Prediction is performed employing a mutation database obtained from UniProtKB/Swiss-Prot and other experimental data previously generated from mutagenesis experiments [18].
- MutationAssessor predicts the functional impact of amino acid substitutions on proteins using the evolutionary conservation of the affected amino acid in protein counterparts. Multiple Sequence Alignment is used to reflect functional specificity, represent the functional impact of a missense variant, and generate conservation scores. Variants with higher scores are more likely to be pathogenic [19].
- MutationTaster2 predicts functional changes in DNA sequences. It is designed to predict consequences based on amino acid substitutions, and intronic substitutions such as synonymous changes, short insertion or exclusion mutations, and variants that cover the limits of intron and exons [20].
- Likelihood Ratio Test (LRT) is a metric that evaluates the proportion of synonymous and non-synonymous mutations in protein-coding regions. When the proportion of mutations is irregular it means that a negative selection process occurred over that region during evolution, which consequently modifies codons in peptide chains [21].

2.3. Data pre-processing and label encoding

Following functional annotation, the ClinVar dataset was preprocessed using in-house scripts for data extraction, and encoding methods implemented in *scikit-learn*. For this study, we investigate three data encoding strategies, namely a) label encode, b) ordinal with Min-Max scaler, and c) One-hot encoding that transforms categorical variables using a dummy strategy. In this last case, a new variable is created to hold binary values, for each column category.

2.4. Classification methods and model evaluation

The data encoding strategies were combined with three supervised machine learning methods: a) Naive Bayes, b) Random Forest, and c) Support Vector Machines.

- **Naive Bayes (NB)** is a probabilistic classifier algorithm based on Bayes' rule that makes a conditional independence assumption between the predictor variables, given an outcome [22]. When applying Bayes' theorem we analyze the probability of an event (A) occurring, given that another event has already accurred (B). The method concedes that the predictors involved are independent, so one predictor does not influence the other. Bayes' theorem can be calculated as follows:

$$P(\mathbf{A}|\mathbf{B}) = \frac{P(\mathbf{B}|\mathbf{A})P(\mathbf{A})}{P(\mathbf{B})} \tag{1}$$

  Naive Bayes is a simplistic classifier that has been reported to have an excellent performance regarding execution time and accuracy, even when considering analyses of large datasets [23].
- **Random Forest (RF)** was created to circumvent the limitations of a single decision tree. RF is an ensemble classification method that combines a set of fitted trees for classification or regression problems [24]. Essentially, each tree is built using a subset of data by bootstrapping the original set of samples, and also a random subset of features [25]. This random sampling provides a low correlation between individual trees avoiding over-fitting. The final decision is based on averaging the probabilistic prediction for each class, instead of majority vote.
- **Support Vector Machine (SVM)** is a learning machine that can be used in regression or classification problems. It uses a constructive universal learning procedure based on statistical

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 November 2020 doi:10.20944/preprints202011.0519.v1

Clinical impact	Original dataset	Training dataset	Reclassification dataset
Benign	263.742	18.891	-
Pathogenic	124.299	16.471	-
Conflict interpretation	41.634	-	7.560
Uncertain Significance - VUS	345.246	-	100.974

Table 1. Distribution of genetic variants in ClinVar original dataset (September 27, 2020). Training dataset comprises variants fully annotated for eight functional predictors as well as the reclassification dataset

learning theory [26]. SVM aims to find the hyperplane that best separates the classes [27] and can find a nonlinear mapping of high-dimensional data, determining optimal hyperplanes. SVM leads to the nonlinear separation of data in the input space using functions called kernels.

For model evaluation, we calculated the average for accuracy, F1-score, Receiver Operating Curve (ROC), and Area Under Curve (AUC) taking the achievement of 10-fold cross-validation into consideration. All models and metrics were implemented using *scikit-learn* packages.

3. Results

3.1. Selection of meta-prediction model

In order to build the training database, we filtered ClinVar's original data aiming to remove variants with missing data, leaving only variants that were classified by all eight single predictors. The filtering strategy for ClinVar's database resulted in a slightly unbalanced training dataset without missing data, since we observed that more benign variants were catalogued than pathogenic genetic variants. The final set of variants in this process is shown in Table 1, which highlights the significant decrease in the number of genetic variants with completed predictor annotations.

All data were encoded using the three aforementioned strategies, Label encoding, Min-Max Scalar, and One-hot encoding. Such strategies were combined with NB, RF, and SVM, and nine binary classification models were produced. All binary classification models were evaluated by performing 10-fold cross-validation. The average accuracy and F1-score for all models show a similar model performance (see Table 2). The whole distribution of accuracy and F1-score showed low variance values, namely 2e-4 and 5e-3, respectively.

	Label Encoding		Min-Max Scaler		One-hot Encoding	
	Accuracy	**F1-Score**	**Accuracy**	**F1-Score**	**Accuracy**	**F1-Score**
Naive Bayes	0.827	0.828	0.837	0.832	0.844	0.840
Random Forest	0.861	0.851	0.861	0.851	0.861	0.721
SVM - SVC	0.856	0.848	0.856	0.844	0.860	0.850

Table 2. Average accuracy and average F1-Score based on 10-fold cross validation.

We choose the best RF model based on F1-Score and AUC. RF accuracy remained at the same level for all encoding strategies (accuracy = 0.86). Acknowledging that our training dataset was unbalanced (had a higher number of benign variants than pathogenic), RF showed better performance than NB and SVM on predicting pathogenicity, except when data were one-hot encoded (F1-Score = 0.72). In a ranking through AUC, RF models achieved better and more consistent predictions for all encoding data strategies with AUC = 0.93. Figure 1 shows ROC and AUC for all nine models.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 November 2020 doi:10.20944/preprints202011.0519.v1

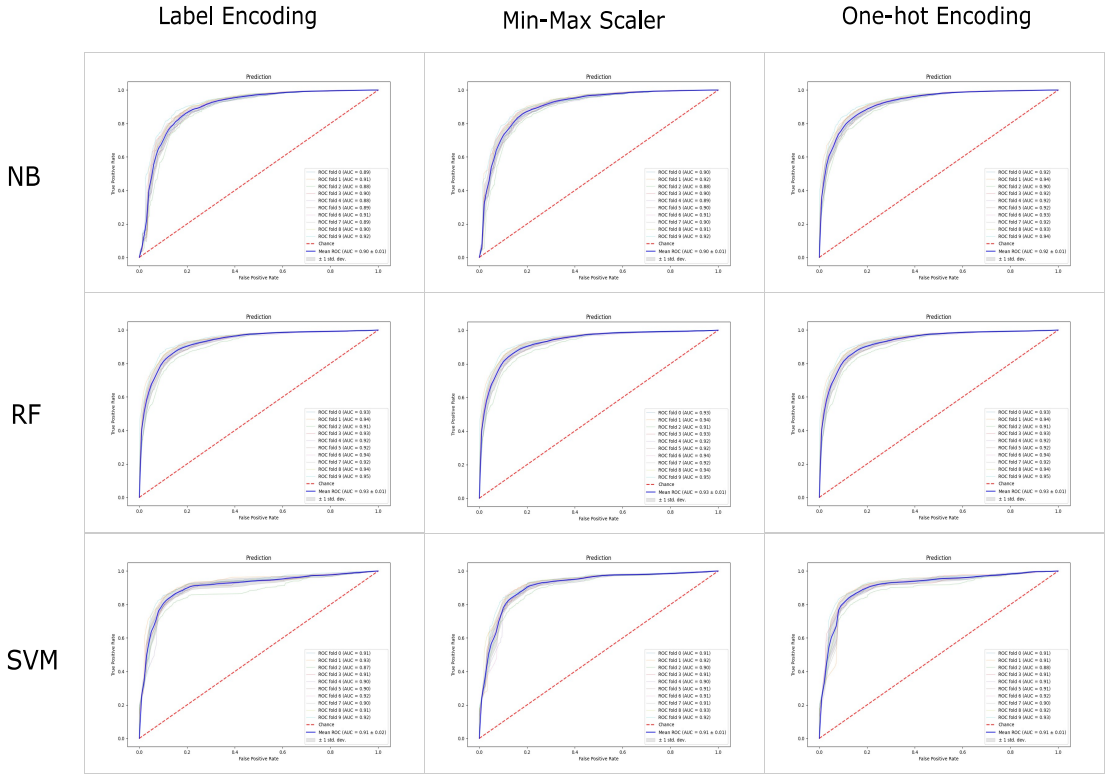

Figure 1. Results of 10-cross validation and ROC plots for all models Naive Bayes (NB) models, Random Forest (RF), and Support Vector Machine (SVM) considering each data encoding method.

3.2. Reclassification of VUS and CI variants

We applied the RF model trained with Label Encoding to reclassify 108,534 genetic variants. As a result, 34,522 (31.8%) VUS and 2,468 (2.27%) CI variants were labeled as pathogenic variants. The distribution of all VUS and CI variants was specified as indicated in Table 3. For a deeper variant analysis, we investigated the RF pathogenic probability distribution for VUS and CI variants (in Figure 2).

Previous Classification	Reclassification	Count	%
VUS	Benign	66,452	65.81
	Pathogenic	34,522	34.19
CI	Benign	5,092	67.35
	Pathogenic	2,468	32.65

Table 3. Counts of benign/pathogenic variants after reclassification by Random Forest model.

We identified a set of 5,297 genetic variants with RF high-probability of pathogenicity according to Random Forest analyses(RF_{prob}>=0.9) 2. These variants were distributed throughout 1,019 gene regions. Reactome pathway analysis was performed for those genes, which revealed a set of 24 enriched pathways (see Table 4).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 November 2020 doi:10.20944/preprints202011.0519.v1

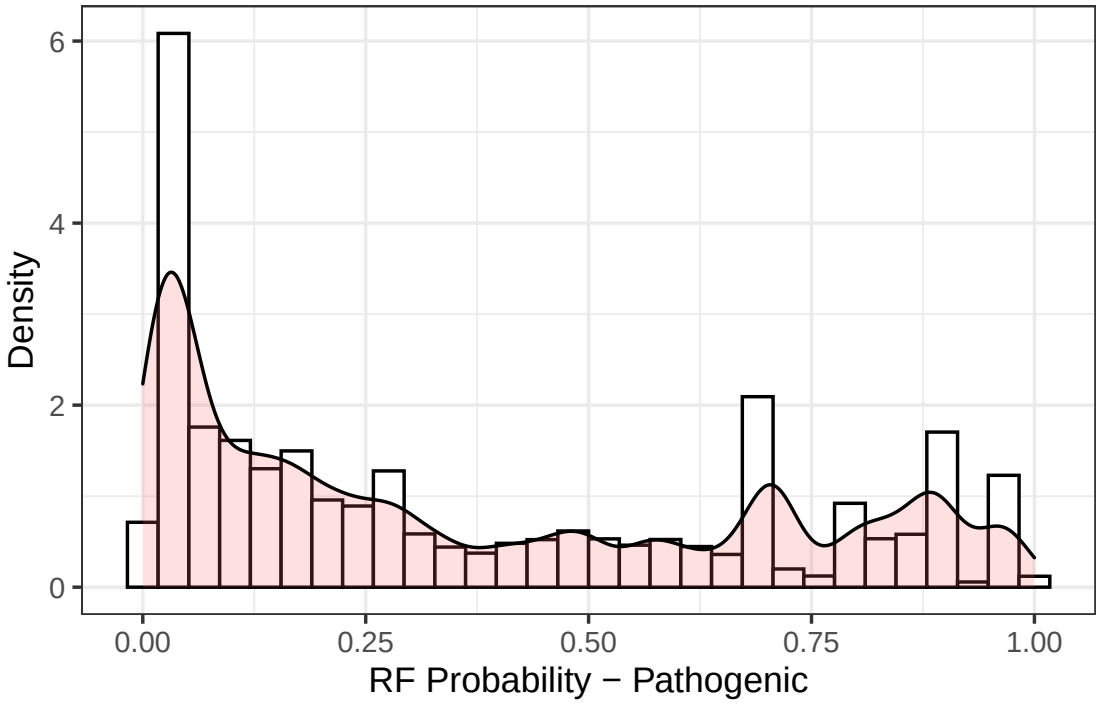

Figure 2. Histogram and density of pathogenicity probability predicted by Random Forest model.

Table 4. Pathway enrichment of genes mapped for VUS and CI genetic variants with 0.9 pathogenicity probability.

Reactome Pathway	Number of Genes	FDR
Diseases of metabolism	76	3.19E-07
Extracellular matrix organization	56	1.40E-04
Muscle contraction	42	0.00200118
Degradation of the extracellular matrix	37	4.41E-06
Diseases of glycosylation	35	0.0048245
Cardiac conduction	30	0.00821261
Signaling by PDGF	26	1.15E-05
ECM proteoglycans	25	1.07E-05
Integrin cell surface interactions	25	4.56E-05
Collagen formation	25	4.92E-04
Collagen degradation	24	4.41E-06
Signaling by MET	23	3.34E-04
Collagen biosynthesis and modifying enzymes	22	1.40E-04
Assembly of collagen fibrils and other multimeric structures	21	9.67E-05
NCAM signaling for neurite out-growth	21	1.40E-04
Collagen chain trimerization	20	1.97E-06
Non-integrin membrane-ECM interactions	19	2.27E-04
NCAM1 interactions	16	2.27E-04
MET promotes cell motility	14	0.0035418
MET activates PTK2 signaling	13	5.12E-04
Laminin interactions	12	0.00177183
Signaling by PDGFRA extracellular domain mutants	9	0.00355751
Signaling by PDGFRA transmembrane, juxtamembrane and kinase domain mutants	9	0.00355751
Anchoring fibril formation	8	0.00399627

The enriched pathways are associated with many important cell functions, such as metabolic processes, cell growth and division, extracellular matrix organization and degradation, muscle

 doi:10.20944/preprints202011.0519.v1

contraction and cardiac conduction. Thus, missense variants related to these pathways may disrupt biological process activity.

3.3. AmazonForest: web platform for variant classification

We developed the AmazonForest to improve user experience towards pathogenicity prediction. AmazonForest was implemented as an online platform that performs our best meta-prediction model to predict pathogenicity of VUS, CI, and new genetic variants. AmazonForest can be accessed at https://www2.lghm.ufpa.br/amazonforest. The platform is divided into two components:

- **The user interface component.** AmazonForest was developed as a web tool with an interface that allows performing pathogenicity prediction of SNPs or INDELs with *in silico* analysis employing the aforementioned best meta-predictor model. The simple web interface enables the user to predict pathogenicity in two ways: first, by providing genomic or dbSNP information (chromosome, chromosome position, or rsID), and second, the user can combine predictor's results to query pathogenicity status (v4.3).
- **Model administrator component.** We implemented a model administrator component to further assess the evolution and performance of the model. This model component enables reproducibility of up-to-date data.

We detailed the business process model and notation as well as use cases in Supplementary Material Section 1. In Supplementary Material Section 2, we show a usage example of AmazonForest's web interface for querying and meta-model prediction of pathogenicity. The web module was developed using Python [28], Javascript (https://developer.mozilla.org/en-US/docs/Web/JavaScript/Reference), HTML5 (https://developer.mozilla.org/pt-BR/docs/Web/HTML/HTML5), and using frameworks such as Flask (https://palletsprojects.com/p/flask/), scikit-learn [29], Pandas [30], Numpy [31].

4. Discussion

In this study, we evaluated the performance of data pre-processing methods (Label Encoding, Min-Max Scalar and One-hot encoding) combined with classical prediction methods, such as Naive Bayes, Random Forest, and Support Vector Machine to produce a meta-prediction model (AmazonTree). The best RF model was adopted for classification of VUS and CI variants.

Some previously proposed meta-prediction approaches were proposed each containing different machine learning or statistical methods, and differ in training datasets [9–11,13]. Futher, most of the reviewed meta-predictors used decision tree methods, [9,11,13]. Decision tree-based models deal with categorical predictors without the need to transform them [32]. However, they are unclear about how they handle missing data, which may produce biased models. To avoid data bias and to obtain a reliable and robust model, our study excluded variants with missing data from the training set, and also VUS and CI variants were also classified if they showed data for the aforementioned eight predictors.

This study is the first to investigates different encoding methods and their influence on pathogenic predictions by NB, RF, and SVM. We found that encoding methods had little influence on models (see ROC and AUC in Figure 1) and that RF is more robust than NB and SVM, except in the case when one-hot strategy was used for data encoding [32].

Accuracy is the predominant metric in meta-prediction studies. Despite being widely employed, accuracy parameters are not enough to evaluate model performance. Global accuracy may be predisposed to bias in training or test data. Therefore, other evaluation metrics are combined for a solid decision and model adoption [33,34].

To the best of our knowledge, there are no similar works that offer or provide tools for user interaction with the user. [9–11,13]. Thus, here we have provided an online tool for a better understanding of the training process and variants reclassification. The tool also allows the

visualization of each steps and results of applying the selected models. In addition, the AmazonForest toold can be applied for new adaptations of other models and their results.

Our proposed model was used for pathogenicity prediction of VUS and CI variants. After prediction, we identified a valuable set of 5,297 variants, at an RF probability cut-off >= 0.9. This cut-off produces a variant set with a high-probability of being pathogenic. Pathway enrichment analysis revealed that sets of genes mapped for new pathogenic variants (VUS and CI), were associated with crucial pathways, that may play important functions at a cellular level. This information could further improve our understanding of well-known diseases, as well as clarify molecular mechanisms involved in rare disorders. Therefore, our tool can help to conduct more careful and accurate analyses of variants of uncertain significance and CI.

5. Conclusion

Finally, our benchmark shows that single machine learning (Naive Bayes and SVM) and ensemble methods present satisfactory prediction results (AUC > 0.91) regarding categorical data and encoding strategies for eight functional predictors. Our results also revealed that when comparing these three machine learning approaches, Random Forests obtained better results. Here, we provide a new web database for the reclassification of a large set of VUS and CI variants. Clinicians may consider this new genetic variant data and web tool to assist their decision process during treatments, while geneticists could research the pathogenicity of diseases in the study of population genomics. Future studies will include research on how model performance is affected by missing data.

Author Contributions: HP handled data, built RF models, and implemented the AmazonForest online platform. Ândrea Ribeiro-dos-Santos and GSA designed the study. All authors wrote and collaboratively reviewed the manuscript.

Funding: This research was funded by and PROPESP/UFPA for the financial support and scholarships.

Acknowledgments: We would like to thank reviewers from Rede de Pesquisa em Genômica Populacional Humana/Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES (Bio. Computacional, No. 3381/2013); Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

NGS	Next Generation Sequencing
GWAS	Genome Wide Association Studies
SNP	Single Nucleotide Polymorphism
ML	Machine Learning
VUS	Variants of uncertain significance
CI	Conflicting interpretations
NB	Naive Bayes
RF	Random Forest
SVM	Support Vector Machine
ROC	Receiver Operating Curve
AUC	Area Under Curve

References

1. MacArthur, J.; Bowler, E.; Cerezo, M.; Gil, L.; Hall, P.; Hastings, E.; Junkins, H.; McMahon, A.; Milano, A.; Morales, J.; others. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic acids research* **2017**, *45*, D896–D901.
2. Araújo, G.S.; Lima, L.H.C.; Schneider, S.; Leal, T.P.; da Silva, A.P.C.; Vaz de Melo, P.O.; Tarazona-Santos, E.; Scliar, M.O.; Rodrigues, M.R. Integrating, summarizing and visualizing GWAS-hits and human diversity with DANCE (Disease-ANCEstry networks). *Bioinformatics* **2016**, *32*, 1247–1249.

3. Deng, N.; Zhou, H.; Fan, H.; Yuan, Y. Single nucleotide polymorphisms and cancer susceptibility. *Oncotarget* **2017**, *8*, 110635.
4. Unoki, H.; Takahashi, A.; Kawaguchi, T.; Hara, K.; Horikoshi, M.; Andersen, G.; Ng, D.P.; Holmkvist, J.; Borch-Johnsen, K.; Jørgensen, T.; others. SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations. *Nature genetics* **2008**, *40*, 1098–1102.
5. Araújo, G.S.; Souza, M.R.; Oliveira, J.R.M.; Costa, I.G. Random Forest and Gene Networks for Association of SNPs to Alzheimer's Disease. Brazilian Symposium on Bioinformatics. Springer, 2013, pp. 104–115.
6. Souza, M.; Araujo, G.; Costa, I.; Oliveira, J.; Initiative, A.D.N.; others. Combined genome-wide CSF Aβ-42's associations and simple network properties highlight new risk factors for Alzheimer's disease. *Journal of Molecular Neuroscience* **2016**, *58*, 120–128.
7. Landrum, M.J.; Lee, J.M.; Riley, G.R.; Jang, W.; Rubinstein, W.S.; Church, D.M.; Maglott, D.R. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic acids research* **2013**, *42*, D980–D985.
8. Alzu'bi, A.A.; Zhou, L.; Watzlaf, V.J. Genetic variations and precision medicine. *Perspectives in health information management* **2019**, *16*.
9. Ranganathan Ganakammal, S.; Alexov, E. An Ensemble Approach to Predict the Pathogenicity of Synonymous Variants. *Genes* **2020**, *11*, 1102.
10. Hassan, M.S.; Shaalan, A.; Dessouky, M.; Abdelnaiem, A.E.; ElHefnawi, M. Evaluation of computational techniques for predicting non-synonymous single nucleotide variants pathogenicity. *Genomics* **2019**, *111*, 869–882.
11. Jaravine, V.; Balmford, J.; Metzger, P.; Boerries, M.; Binder, H.; Boeker, M. Annotation of Human Exome Gene Variants with Consensus Pathogenicity. *Genes* **2020**, *11*, 1076.
12. Dong, C.; Wei, P.; Jian, X.; Gibbs, R.; Boerwinkle, E.; Wang, K.; Liu, X. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Human molecular genetics* **2015**, *24*, 2125–2137.
13. do Nascimento, P.M.; Medeiros, I.G.; Falcão, R.M.; Stransky, B.; de Souza, J.E.S. A decision tree to improve identification of pathogenic mutations in clinical practice. *BMC medical informatics and decision making* **2020**, *20*, 1–11.
14. Richards, S.; Aziz, N.; Bale, S.; Bick, D.; Das, S.; Gastier-Foster, J.; Grody, W.; Hegde, M.; Lyon, E.; Spector, E.; Voelkerding, K.; Rehm, H. 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* **2015**, *17*, 405–424. doi:10.1038/gim.2015.30.
15. Shihab, H.A.; Gough, J.; Cooper, D.N.; Stenson, P.D.; Barker, G.L.; Edwards, K.J.; Day, I.N.; Gaunt, T.R. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Human mutation* **2013**, *34*, 57–65.
16. Kumar, P.; Henikoff, S.; Ng, P.C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature protocols* **2009**, *4*, 1073.
17. Adzhubei, I.; Jordan, D.M.; Sunyaev, S.R. Predicting functional effect of human missense mutations using PolyPhen-2. *Current protocols in human genetics* **2013**, *76*, 7–20.
18. Choi, Y.; Chan, A.P. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* **2015**, *31*, 2745–2747.
19. Reva, B.; Antipin, Y.; Sander, C. Determinants of protein function revealed by combinatorial entropy optimization. *Genome biology* **2007**, *8*, R232.
20. Schwarz, J.M.; Cooper, D.N.; Schuelke, M.; Seelow, D. MutationTaster2: mutation prediction for the deep-sequencing age. *Nature methods* **2014**, *11*, 361–362.
21. Chun, S.; Fay, J.C. Identification of deleterious mutations within three human genomes. *Genome research* **2009**, *19*, 1553–1561.
22. Mitchell, T.M. Generative and discriminative classifiers: Naive bayes and logistic regression. *Machine learning* **2010**, pp. 1–17.
23. Rish, I.; others. An empirical study of the naive Bayes classifier. IJCAI 2001 workshop on empirical methods in artificial intelligence, 2001, Vol. 3, pp. 41–46.
24. Breiman, L. Random forests. *Machine learning* **2001**, *45*, 5–32.

25. Strobl, C.; Boulesteix, A.L.; Zeileis, A.; Hothorn, T. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics* **2007**, *8*, 25.
26. Cortes, C.; Vapnik, V. Support-vector networks. *Machine learning* **1995**, *20*, 273–297.
27. Xue, H.; Yang, Q.; Chen, S. SVM: Support vector machines. *The top ten algorithms in data mining* **2009**, *6*, 37–60.
28. Van Rossum, G.; Drake, F.L. *Python 3 Reference Manual*; CreateSpace: Scotts Valley, CA, 2009.
29. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.
30. Wes McKinney. Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference; Stéfan van der Walt.; Jarrod Millman., Eds., 2010, pp. 56 – 61. doi:10.25080/Majora-92bf1922-00a.
31. Harris, C.R.; Millman, K.J.; van der Walt, S.J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N.J.; Kern, R.; Picus, M.; Hoyer, S.; van Kerkwijk, M.H.; Brett, M.; Haldane, A.; del R'ıo, J.F.; Wiebe, M.; Peterson, P.; G'erard-Marchant, P.; Sheppard, K.; Reddy, T.; Weckesser, W.; Abbasi, H.; Gohlke, C.; Oliphant, T.E. Array programming with NumPy. *Nature* **2020**, *585*, 357–362. doi:10.1038/s41586-020-2649-2.
32. Au, T.C. Random forests, decision trees, and categorical predictors: the" absent levels" problem. *The Journal of Machine Learning Research* **2018**, *19*, 1737–1766.
33. Gonçalves, W.G.; Ribeiro, H.M.C.; Sá, J.A.S.d.; Morales, G.P.; Ferreira Filho, H.R.; Almeida, A.d.C. Classification of forest types using artificial neural networks and remote sensing data. *Revista Ambiente & Água* **2016**, *11*, 612–624.
34. e Gonçalves, W.G.; dos Santos, M.H.d.P.; Lobato, F.M.F.; Ribeiro-dos Santos, Â.; de Araújo, G.S. Deep learning in gastric tissue diseases: a systematic review. *BMJ Open Gastroenterology* **2020**, *7*, e000371.