VaRTK – An accurate machine learning model trained on clinically curated variants to predict the variant pathogenicity score

Ravi Gupta, Manju Lakshmi, Charugulla Sai Yuva Sandeep, S. G. Thenral, Sandhya Nair, Anurag Gupta, Thiramsetti Sattibabu, Amit Parhar, Tamanna Golani, Sudarshana J. Pai, Tavleen Bajwa, Sakthivel Murugan S. M., Ramprasad V. L.

MedGenome Labs, 3rd Floor, Narayana Netralaya Building Narayana Health City, #258/A, Bommasandra, Hosur Rd, Bengaluru, Karnataka 560099, India
Email – ravig@medgenome.com

Abstract

Identifying and prioritizing the disease causing variants from thousands of variants that are called during a whole genome and exome sequencing analysis is a time consuming and manual task. Pathogenicity based ranking of variants greatly improves the speed of report generation, in turn increasing the diagnostic yield.

Machine learning models recognize patterns in variant data and help in informed decision making and thus reduce time in the variant prioritization process. Models trained on public disease databases have not been so successful because of the presence of noisy and unvalidated data.

We have developed a supervised machine learning model trained on manually curated and reviewed disease causing variants from a cohort of ~100,000 clinical samples. The VaRTK (**Va**riant **R**anking **T**ool**K**it) model is a random forest model trained on 46,345 (10,336 disease causing and 36,009 benign variants) which cover ~10,000 genes.

Our model was able to achieve a F1 and sensitivity of 0.98 on unseen test data consisting of ~15,000 variants covering ~6,000 genes. The model performance was further evaluated on a validation cohort of 166 samples which were resolved through manual interpretation. The model ranked the manually selected pathogenic variant(s) in the Top 20 in 90.66% of cases.

Model Building

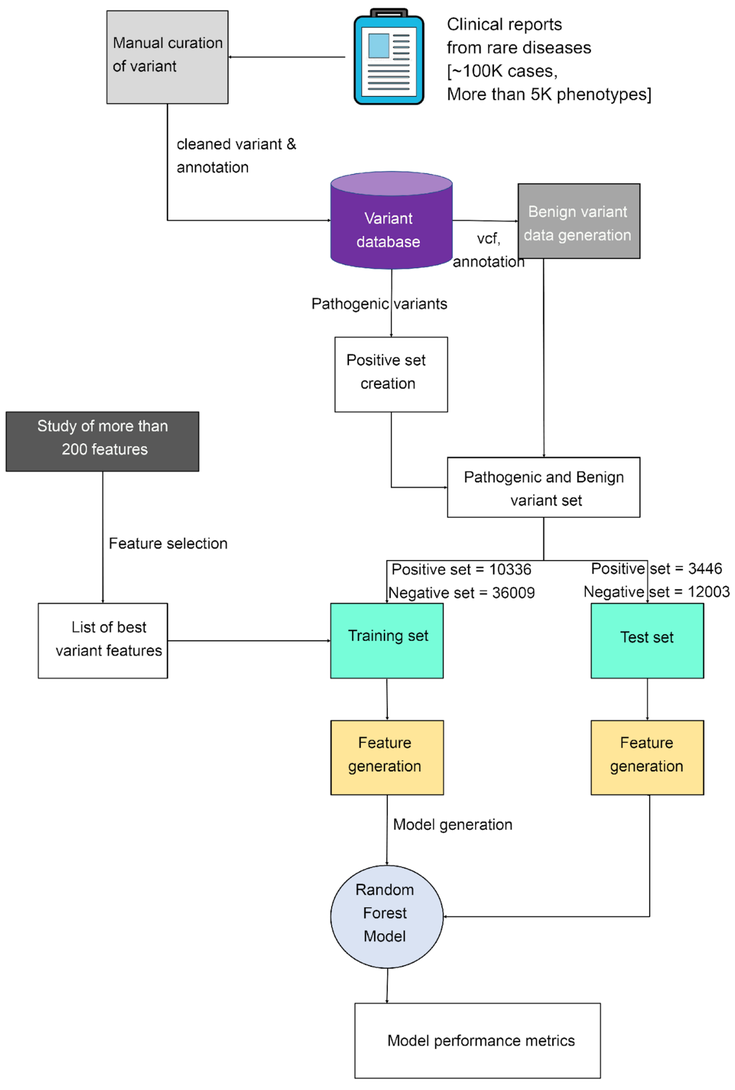

Figure 1. VaRTK model development
The figure illustrates the model development process which involves curating unique training set from our in-house variant database which stores all pathogenic, likely pathogenic and variants of uncertain significance reported in the clinical diagnosis of a sample. We have built the model from more than 100,000 clinical case reports. The positive set is derived from pathogenic and likely pathogenic variants reported in the clinical reports. The negative set (benign) is derived from the pathogenic/likely pathogenic case reports. Variants are randomly sampled from the pathogenic/likely pathogenic reports after removing the pathogenic variant in the report. We performed a careful random sampling to balance genes, variant class and variant allele frequency. We also performed study of more than 200 variant features and identified a set of 36 features which are used for model building. A well studied list of features are extracted for each of the variants which incorporates in-silico tool prediction scores, allele frequencies, disease databases, internal frequency databases and few variant level identifiers. The dataset is then split into train and test. The training set is used as the input for the random forest model. The model gives us a prediction score between 0 to 1, the higher scores denoting deleterious nature of the variant.

VaRTK Dataset

Table 1. Train and test dataset
We trained four models to see which performs better when encountered with class imbalances and variant class. The first two models are trained on all variant classes while the third and fourth model are only trained on missense variants.

Model	Train Positive	Train Negative	Test Positive	Test Negative
Unbalanced dataset	10,336	36,009	3,446	12,003
Balanced dataset	10,336	10,336	3,446	12,003
Missense unbalanced	3,932	32,553	3,446	12,003
Missense un-balanced	3,932	3,932	3,446	12,003

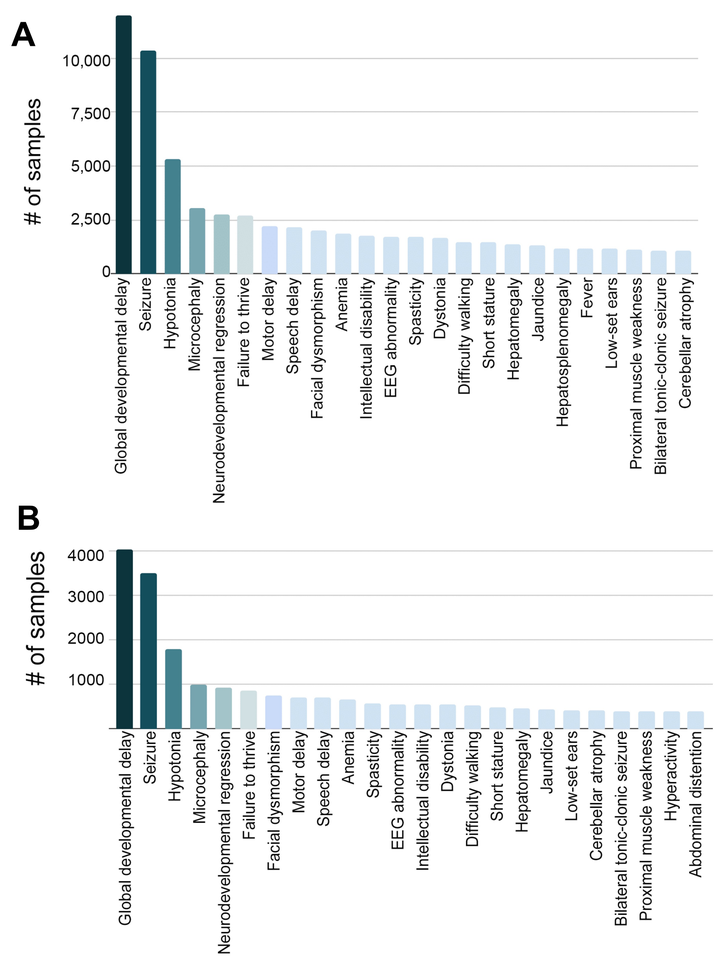

Figure 2. Train and test set top 20 HPO phenotypes
The top 20 HPO terms observed in the training **(A)** and test **(B)** set samples is provided here. Since a large number of samples that we receive is related to neurology; global development delay, seizure, hypotonia are the top phenotype terms observed in our training set.

VaRTK Model Performance

Table 2. Performance metrics of model on unseen test set

Model	F1	Accuracy	Precision	Recall	AUC
Unbalanced dataset	0.98	0.99	0.99	0.98	0.999
Balanced dataset	0.98	0.99	0.97	0.99	0.999
Missense unbalanced	0.95	0.98	0.99	0.91	0.999
Missense un-balanced	0.96	0.98	0.95	0.97	0.998

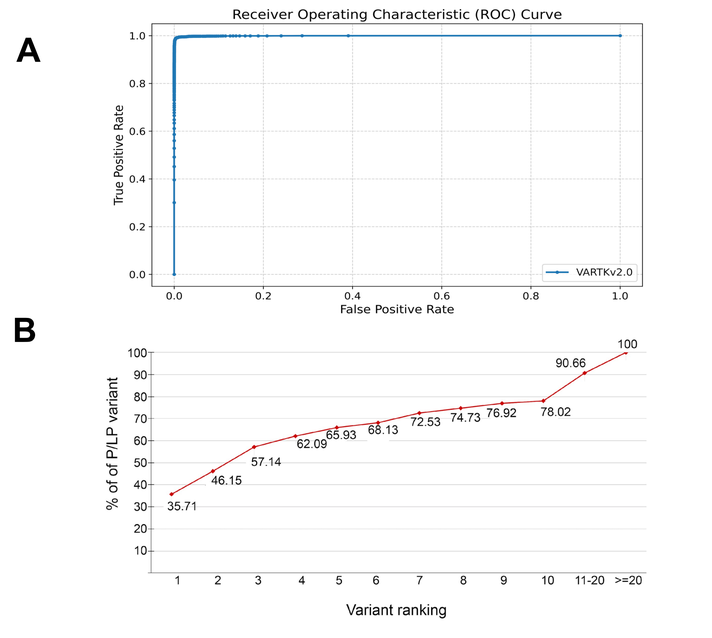

Figure 3. Performance metrics
(A) ROC on the test using unbalanced model. **(B)** Performance of VaRTK model on 166 clinical cases which were manually resolved. The model was able to capture the causative variant in the top 20 ranks in 90.66% of the cases without any gene based filter applied on the variants.

Comparison with other *in-silico* tools

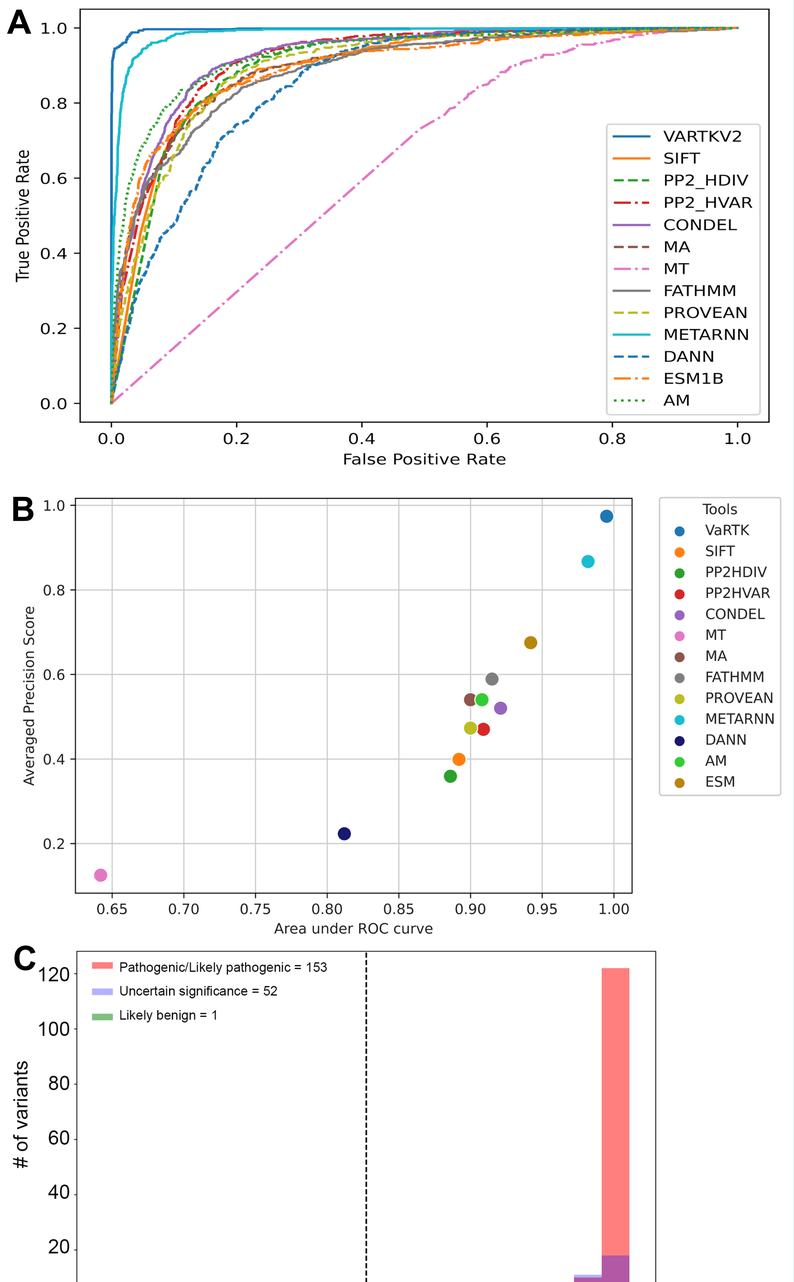

Figure 4. Comparison of VaRTK model with existing *in silico* tools
(A) VaRTK model was compared against 12 other nsSNV prediction tools (MA – Mutation Assessor, MT – Mutation taster, AM – AlphaMissense), using unseen test set of ~2,600 clinically reported rare missense variants. The variants were selected based on their absence in the ClinVar database to reduce circularity of data as it has been observed that some of the prediction tools use the variants from ClinVar database for training. **(B)** Comparison of VaRTK model performance against other *in-silico* prediction tools using the same test as mentioned in **(A)**. The AP (average precision) versus AUC plot is an important measure of evaluating a model performance when the train set classes are unbalanced. Average precision indicates whether a model can correctly identify all the positive examples with low amount of false positives. VaRTK shows better performance across both metrics in comparison to other models. **(C)** Analysing the performance of VaRTK model on variants used in recently published study from (Emedgene) Meng et. al. The VaRTK model was able to classify 98% of the variants into the correct corresponding category.

Summary

- ❖ VaRTK is trained on large number of manually curated diverse clinically classified variants. The benign set is also picked systematically to obtain a training set unbiased by specific features.
- ❖ A set of 36 features are selected from 200+ variant features that provides the best performance.
- ❖ VaRTK model achieves an F1 and Recall score of 0.98 with AUC as 0.99.
- ❖ VaRTK model is able to rank 90.7% of the manually selected variants in top 20 among all the variants in the exomes.
- ❖ Comparison with other in-silico model shows that VaRTK achieves better performance metrics over other tools including Google AlphaMissense.

References

- ❖ Li, C., Zhi, D., Wang, K. *et al.* MetaRNN: differentiating rare pathogenic and rare benign missense SNVs and InDels using deep learning. *Genome Med* **14**, 115 (2022).
- ❖ Jun Cheng et al.,Accurate proteome-wide missense variant effect prediction with AlphaMissense.Science381,eadg7492(2023).
- ❖ Brandes, N., Goldman, G., Wang, C.H. et al. Genome-wide prediction of disease variant effects with a deep protein language model. Nat Genet 55, 1512–1522 (2023).
- ❖ Liu, X., Li, C., Mou, C. et al. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. Genome Med 12, 103 (2020)
- ❖ Meng L, Attali R, Talmy T, Regev Y, Mizrahi N, Smirin-Yosef P, Vossaert L, Taborda C, Santana M, Machol I, Xiao R, Dai H, Eng C, Xia F, Tzur S. Evaluation of an automated genome interpretation model for rare disease routinely used in a clinical genetic laboratory. Genet Med. (2023).