Original Paper

Diagnosis of a Single-Nucleotide Variant in Whole-Exome Sequencing Data for Patients With Inherited Diseases: Machine Learning Study Using Artificial Intelligence Variant Prioritization

Yu-Shan Huang[1], MSc; Ching Hsu[2], MSc; Yu-Chang Chune[1], MSc; I-Cheng Liao[1], MSc; Hsin Wang[2], MSc; Yi-Lin Lin[3], MSc; Wuh-Liang Hwu[4], MD, PhD; Ni-Chung Lee[3], MD, PhD; Feipei Lai[1,2], PhD

[1]Department of Computer Science and Information Engineering, National Taiwan University, Taipei City, Taiwan

[2]Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei City, Taiwan

[3]Department of Medical Genetics, National Taiwan University Hospital, Taipei City, Taiwan

[4]Department of Pediatrics, National Taiwan University Hospital, Taipei City, Taiwan

Corresponding Author:
Feipei Lai, PhD
Graduate Institute of Biomedical Electronics and Bioinformatics
National Taiwan University
Number 1, Roosevelt Road, Section 4
Taipei City, 106319
Taiwan
Phone: 886 2 33664924
Fax: 886 2 23628167
Email: flai@ntu.edu.tw

Abstract

Background: In recent years, thanks to the rapid development of next-generation sequencing (NGS) technology, an entire human genome can be sequenced in a short period. As a result, NGS technology is now being widely introduced into clinical diagnosis practice, especially for diagnosis of hereditary disorders. Although the exome data of single-nucleotide variant (SNV) can be generated using these approaches, processing the DNA sequence data of a patient requires multiple tools and complex bioinformatics pipelines.

Objective: This study aims to assist physicians to automatically interpret the genetic variation information generated by NGS in a short period. To determine the true causal variants of a patient with genetic disease, currently, physicians often need to view numerous features on every variant manually and search for literature in different databases to understand the effect of genetic variation.

Methods: We constructed a machine learning model for predicting disease-causing variants in exome data. We collected sequencing data from whole-exome sequencing (WES) and gene panel as training set, and then integrated variant annotations from multiple genetic databases for model training. The model built ranked SNVs and output the most possible disease-causing candidates. For model testing, we collected WES data from 108 patients with rare genetic disorders in National Taiwan University Hospital. We applied sequencing data and phenotypic information automatically extracted by a keyword extraction tool from patient's electronic medical records into our machine learning model.

Results: We succeeded in locating 92.5% (124/134) of the causative variant in the top 10 ranking list among an average of 741 candidate variants per person after filtering. AI Variant Prioritizer was able to assign the target gene to the top rank for around 61.1% (66/108) of the patients, followed by Variant Prioritizer, which assigned it for 44.4% (48/108) of the patients. The cumulative rank result revealed that our AI Variant Prioritizer has the highest accuracy at ranks 1, 5, 10, and 20. It also shows that AI Variant Prioritizer presents better performance than other tools. After adopting the Human Phenotype Ontology (HPO) terms by looking up the databases, the top 10 ranking list can be increased to 93.5% (101/108).

Conclusions: We successfully applied sequencing data from WES and free-text phenotypic information of patient's disease automatically extracted by the keyword extraction tool for model training and testing. By interpreting our model, we identified which features of variants are important. Besides, we achieved a satisfactory result on finding the target variant in our testing data set. After adopting the HPO terms by looking up the databases, the top 10 ranking list can be increased to 93.5% (101/108).

The performance of the model is similar to that of manual analysis, and it has been used to help National Taiwan University Hospital with a genetic diagnosis.

(JMIR Bioinform Biotech 2022;3(1):e37701) doi: 10.2196/37701

KEYWORDS

next-generation sequencing; genetic variation analysis; machine learning; artificial intelligence; whole-exome sequencing

Introduction

Background

Modern next-genome sequencing (NGS) technology makes rapid human genome sequencing within a day possible [1,2]. Because of its speed and low cost in comparison with the traditional Sanger sequencing method [3], NGS is being rapidly introduced into clinical and public health laboratory practice, especially for the diagnosis of hereditary disorders.

Although NGS has extremely high throughput and could generate huge amounts of genomic data in a short time, interpreting these data and finding the disease-causing candidates among thousands of variants remain a challenge. To determine the true causal variants of a patient with genetic disease, physicians often need to view numerous features on every variant manually and search for literature in different databases to understand the effect of a genetic variation. Another challenge is in finding the genetic variants that have a strong correlation with patient's phenotype. Physicians often select useful keywords from patient's electronic medical records (EMRs) manually to search for articles in several genetic databases such as Online Mendelian Inheritance in Man (OMIM) [4] and GeneReviews [5] to decide whether a variant is correlated with a genetic disease. It is thus a burden for physicians to go through these laborious and time-consuming processes case-by-case, especially when the number of inherited disease–associated germline mutations published per year has increased exponentially in the last decade [6].

Nowadays, many studies use machine learning methods to solve numerous problems in genomics and genetics. The field of machine learning promises to enable computers to assist humans in making sense of large, complex data sets. After variant annotation, there is a variant list with hundreds of columns that humans are not capable of interpreting one-by-one. As machine learning significantly surpasses human-level performance, especially with structured data, we consider using a machine learning method to analyze variants from NGS and find the target gene.

To address these problems, it is important and necessary to have a high-performance method to filter candidate variants from NGS results and immediately find target variants related to a patient's disease. Recently, many tools such as Exomiser [7], DeepPVP [8], Xrare [9], VarSight [10], Phenolyzer [11], Fabric GEM [12], MOON [2], CADD [13], and MetaSVM [14] have been developed to identify potentially causative variants that are relevant to patient's phenotype in rare disease diagnosis. Exomiser integrates information including calculated gene-specific phenotype score, variant allele frequency (Multimedia Appendix 1), and predicted pathogenicity of several alleles to prioritize disease-causative variants/interactions. Fabric GEM utilizes Bayes factor to prioritize variants with the support of a gene-phenotype score calculated by Phevor [15] and variant prioritization result of several tools including ANNOVAR, VAAST, and Phen-Gen. MOON integrates the result of annotation of several variants and prioritization tools to achieve variant prioritization using several kinds of machine learning models. Gene-phenotype scores calculated by Phevor using Human Phenotype Ontology (HPO) terms extracted from electronic health records (EHRs) of patients are also considered by MOON. CADD utilizes logistic regression to integrate information including context of surrounding sequence, biological constraints, epigenetic measurements, and result of several variant annotation tools to build a predictive model for variant deleteriousness. MetaSVM [14] gathers result of 9 deleteriousness prediction scores including PolyPhen-2 [16], SIFT [17], MutationTaster [18] to build a support vector machine (SVM) deleteriousness predictive model. Although these tools adopt different approaches, including logistic regression and deep neural networks, to prioritize variants, most can only recognize the phenotypes defined in the HPO term [19]. In this work, we developed the AI Variant Prioritizer module based on a machine learning approach that can output the rank of single-nucleotide variants (SNVs) and small insertions/deletions (indels) from whole-exome sequencing (WES) data with the input of free-text phenotypic description or EHR.

In this research, we aimed to implement a website, AI Variant Prioritizer, that uses data from NGS bioinformatics pipelines with machine learning to make a prediction about the most possible disease-causing variants among SNVs and patient's phenotype. The data generated from NGS pipelines are all structured with annotations from several tools including ANNOVAR, Nirvana, Variant Effect Predictor (VEP), and InterVar and additional information from multiple databases queried by MViewer (Mutation Viewer) [20]. To simplify the interpretation process, we integrate the keyword extraction tool to generate the phenotype from EMRs automatically. Our system takes candidate variants filtered by MViewer and patient's EMRs as its input and outputs a list of SNVs with rank and probability of being disease causing. Instead of checking every variant manually, this system can assist researchers and physicians in focusing on those with higher disease-causing probability and save a lot of time. Moreover, we implement a web application programming interface (API) for our system so that the ranking function could be integrated into MViewer. Thus, physicians are able to interpret the results of genetic variation with a single application instead of adopting numerous services.

Data Description

In our research, we focus on patients who have been diagnosed with rare Mendelian diseases. Our data are collected mainly from the rapid exome project of Department of Medical Genetics, National Taiwan University Hospital (NTUH). To build the model with more data, we also applied for several WES data that are deposited in the dbGaP database (project ID 20911). The data we use are the dbGaP accession phs000711.v5.p1 by Baylor Hopkins Center for Mendelian Genomics.

The conditions under which we collect patients' sequencing data to meet the requirements of this research are as follows:

- Patients who were diagnosed with genetic disorders.
- Patients who received WES or targeted panel sequencing and diagnosed with at least one disease-causing variant.
- Patients whose phenotype information is available.

Our data from NTUH include patient demographics, variant call format (VCF) file output by the NGS bioinformatics pipeline, and phenotype information from electrical medical records. Data from dbGaP also include patient demographics, VCF file, and clinical conditions. All data are deidentified and will not invade patients' privacy. We include sex in patient demographic information as a feature in our model because some human genetic disorders are sex linked. Sex-linked diseases are caused by mutations in genes on X or Y chromosomes and passed down through families.

Variant Call Format File

As the end product of the NGS bioinformatics pipeline, the VCF is a generic format for storing DNA polymorphism data such as SNPs, insertions, deletions, and structural variants. The format was developed for the 1000 Genomes Project and has also been widely adopted by other projects. Every VCF file consists of 2 two parts: header section and data section. The header contains metadata about the tags and annotations in the data part. It can be also used to provide information related to the history of the data and file. The last line in the header contains the column headings for the data part. The data section is tab separated into 9 columns and reports a mutation for each row. Columns include CHROM, POS, ID, REF, ALT, QUAL, FILTER, INFO, and FORMAT.

Phenotype Information

For the data from NTUH, we extract patient's phenotypic information from clinicians' history summary. It mainly records a brief summary of patient's illness, clinical diagnosis, and the reason(s) why each patient was admitted. We also collect the phenotype keywords provided by doctors based on the symptom of each patient for model validation. For the data from dbGaP, because EHRs are not available, we will use the clinical condition of the patient instead. For the clinical condition that can be found in OMIM databases, we will extract the corresponding description of phenotypes as the phenotypic information to be used in our research.

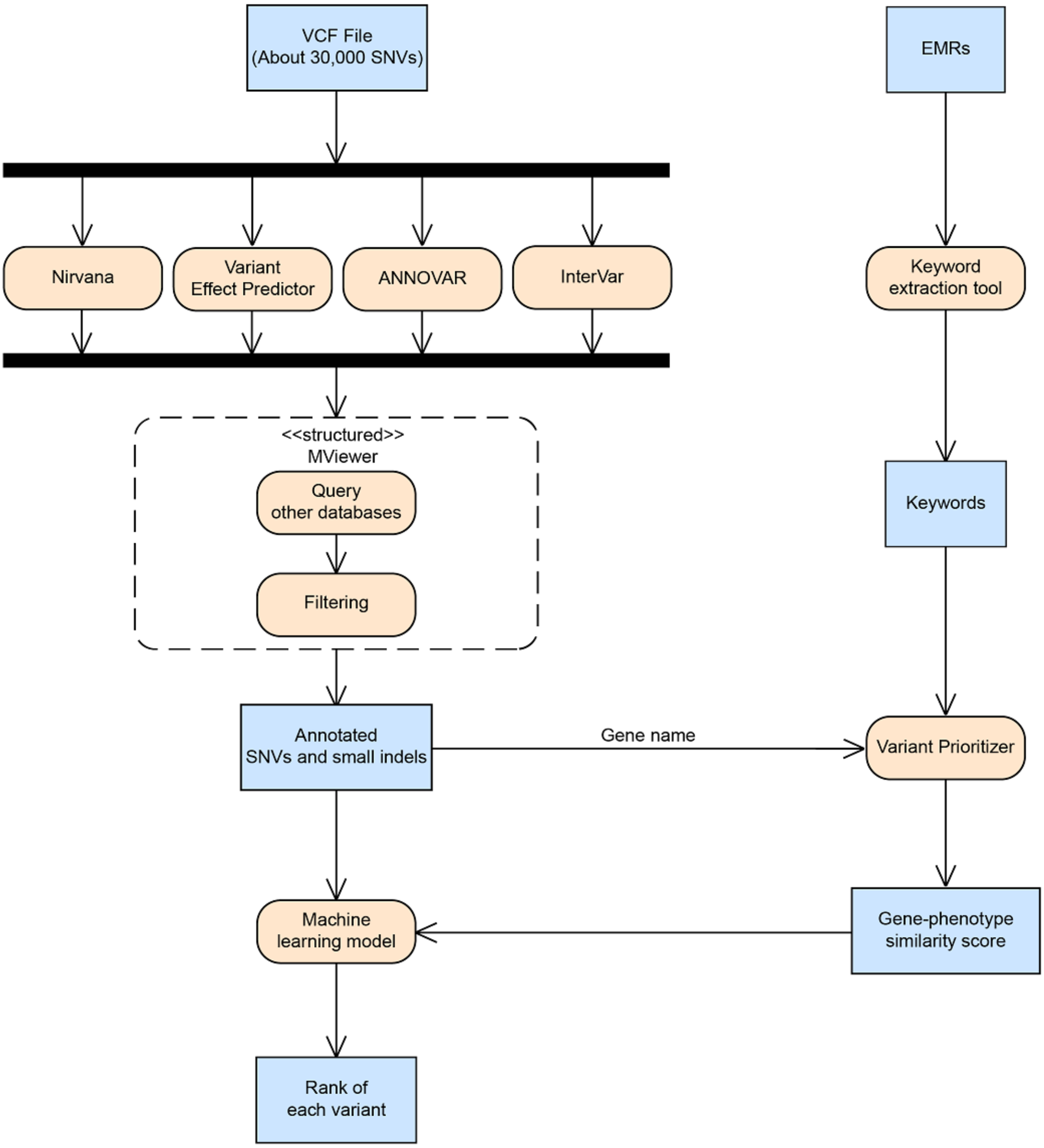
Methods

Workflow

Overview

Figure 1 shows the workflow of our research. We collected VCF of each patient from WES and panel sequencing and then annotated the variants using several tools. After variant annotation, we used our in-house software (MViewer [20]) to query additional external databases and filter for candidate variants. We then used the gene name of these candidate variants and keywords extracted by keyword extraction tools from EMRs to query Variant Prioritizer [21]. The gene similarity scores generated by Variant Prioritizer and columns of annotated variants were used as features to train a machine learning model. This model ranks each variant that represents its disease-causing probability. We will demonstrate the details of each step in the following sections.

Figure 1. The workflow of research. EMR: electronic medical record; indel: insertion/deletion; MViewer: Mutation Viewer; SNV: single-nucleotide variant; VCF: variant call format.



Variant Annotation

We collected each patient's NGS sequencing data in the VCF file and got annotations from several tools, including ANNOVAR [22], VEP [23], Nirvana [24], and InterVar [25]. For additional information that the aforementioned tools will not provide, we used software to import some public data sources, including ClinVar [26], Human Genome Mutation Database (HGMD) [27], and Taiwan Biobank [28]. A detailed description of these annotation fields is summarized in Textbox 1.

Textbox 1. Description of annotation fields.

Allele Frequency

This describes the fraction of gene copies of a particular allele in a defined population. Allele frequency is calculated by dividing the number of copies of a particular allele in a population by the total number of all alleles for that gene in a population. It refers to how common an allele is in a population.

Functional Prediction Score

A range of scoring algorithms with capability to predict the potential deleteriousness of variants based on different information in them, such as their sequence homology, protein structure, and evolutionary conservation. These scoring methods include function prediction scores, conservation scores, and ensemble scores.

Pathogenicity

Clinical significance variants reported in 2 public databases, ClinVar and Human Gene Mutation Database (HGMD), that store information on gene mutation(s) related to human-inherited disease. Both classify variants as disease causing or disease associated by manual curation.

Clinical Interpretation

The American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) published standards and guidelines for the clinical interpretation of sequence variants with respect to human diseases on the basis of 28 criteria [29]. These criteria are as follows: the criteria (16 overall) for classifying variants as pathogenic or likely pathogenic are very strong (PVS1), strong (PS1-PS4), moderate (PM1-PM6), or supporting (PP1-PP5), whereas the criteria (12 overall) for classifying variants as benign or likely benign are standalone (BA1), strong (BS1-BS4), or supporting (BP1-BP7).

Gene-Level Constraint

Constraint on gene expression levels has been shown to influence patterns of genetic variation within humans [30]. For example, some genes are unusually depleted for loss of function and are thought to be constraint with respect to their expression. The Genome Aggregation Database (gnomAD) provides predicted constraint metrics track set that contains metrics of pathogenicity per gene as predicted and identifies genes subject to strong selection against various classes of mutation. These include several subtracks of constraint metrics calculated at gene, transcript, and transcript region levels.

Disease Inheritance

Patterns of inheritance that a trait or disorder associated with a variant can be passed down through families, such as autosomal dominant, autosomal recessive, X-linked, and mitochondrial inheritance. We used the patterns defined in OMIM (Online Mendelian Inheritance in Man) as our data.

Others

Additional information about genetic variants such as the gene name, genotype, and the functional consequence on the different transcripts for a gene or in proximal regulatory regions.

Variant Filtering

There are on average 40,000 variants per proband in WES data. However, most of them are benign and not related to the symptoms. Only a small number of these variants are likely to be deleterious or relevant to the patient’s disease. In a standard clinical analysis process, physicians only focus on variants that might be pathogenic or unknown. As our model aims to assist researchers and physicians with their clinical exome reading, reducing the number of variants and focusing on the variants that are more likely to be responsible for the disease are necessary.

For the purpose of generating candidate variants, we used the filter provided by MViewer to remove the variants that are not likely to be deleterious. The filters and criteria are listed in Table 1. For filters that contain more than 1 column, if a variant meets any of their criterion, it will remain in the data. We got approximately 700 SNVs per patient after variant filtering.

Table 1. Filter criteria.

Filter	Column	Criteria
Max allele frequency	- Max Allele Frequency	- ≤0.01 (include no data)
Nonsynonymous missense mutation	- ExonicFunc.refgene	- "nonsynonymous"
Stop gain	- Consequence - ExonicFunc.refgene	- "stop_gained" - "stopgain"
Splice	- Consequence - Func.refgene	- "splice_region_variant" - "splice_acceptor_variant" - "splice_donor_variant" - "splicing"
Frameshift	- Consequence - ExonicFunc.refgene	- "frameshift_variant" - "feature_truncation" - "feature_elongation" - "frameshift"
Initial codon	- Consequence	- "start_lost"
Deletion	- Type - Consequence - ExonicFunc.refgene	- "deletion"
Insertion	- Type - Consequence - ExonicFunc.refgene	- "insertion"
Inframe deletion	- Consequence - ExonicFunc.refgene	- "inframe_deletion" - "nonframeshift deletion"
Exon/splice site	- Func.refgene - Consequence	- "exonic" - "splicing" - "coding_sequence_variant" - "frameshift_variant" - "incomplete_terminal_codon_variant" - "inframe_deletion" - "inframe_insertion" - "missense_variant" - "splice_acceptor_variant" - "splice_donor_variant" - "splice_region_variant"

Phenotype Extraction

Overview

The phenotype information used in this research is from clinicians' history summary. The records were all free text and the length of texts varied from less than 10 to more than 300 words. In the clinical analysis process, it is time consuming for physicians to go through the medical records and identify the phenotype keywords manually. To solve this problem, we used several keyword extraction tools to automatically generate keywords related to phenotype from free-text medical records. The keyword extraction tools applied in our research are listed in the following sections.

MetaMap

MetaMap [31] is a widely used application providing access to the concepts in the Unified Medical Language System (UMLS) Metathesaurus [32]. The UMLS Metathesaurus is a compilation of names, relationships, and associated information from a variety of biomedical naming systems representing different views of biomedical practice or research. It comprises over 1 million biomedical concepts and 5 million concept names [33]. MetaMap is able to map every word in the texts to UMLS concepts, but we just wanted to focus on those associated with phenotypes and diseases. Thus, we extracted the words that are classified as the semantic types of the following: (1) injury or poisoning, (2) cell or molecular dysfunction, (3) genetic function, (4) disease or syndrome, (5) sign or symptom, (6) tissue.

Doc2Hpo

Doc2Hpo [34] is a web application using natural language processing (NLP) techniques to parse clinical note and get the phenotype concept curation as the HPO term. There is a parsing engine that will automatically recognize the phenotype concepts from the input. Doc2Hpo applies an algorithm called NegBio for negation detection in the input data. After that, there are several NLP engines responsible for HPO concept extraction. We used 3 of these engines and compared the performance of each of them. The first NLP engine is a string-based method that leverages the algorithm for concept extraction. The second engine is the online NCBO Annotator [35] API for HPO concept recognition. The last engine we adopt is MetaMap Lite, which is a fast version of MetaMap that provides near–real-time named entity recognition. The MetaMap Lite engine first identifies clinical terms in the texts and maps them to standard UMLS concepts. The UMLS concepts will then be further mapped to HPO concepts. Results generated by Doc2Hpo are HPO terms, whereas keywords extracted by MetaMap are nonHPO terms.

Phenotype-Gene Similarity Score

Another method to construct the connections between genes and keywords is using the Okapi BM25 ranking function. Okapi BM25 is usually used by search engines, such as Google and Bing, to rank matching documents according to their relevance to a given search. One of the most prominent instantiations of the function is as the following equation:

$$\text{Score}(D, Q) = \sum_{i=1}^{n} \text{IDF}(qi) \cdot f(qi, D) \cdot (k1 + 1)/\{f(qi, D) + k1 \cdot (1 - b + b \cdot |D|/\text{avgdl})\}$$

where score(D, Q) represents the Okapi BM25 score of a document D when given a query Q, containing keywords $q1$, $q2$,...,qn; $f(qi, D)$ is qi’s term frequency in the document D; $|D|$ is the length of document D in words; avgdl is the average document length among all documents; $k1$ and b are constants (=1.2 and 0.8, respectively); and IDF(qi) is the inverse document frequency (IDF) weight of the query term qi and is usually defined as:

$$\text{IDF}(qi) = \ln\ [(N - n(qi) + 0.5]/[n(qi) + 0.5 + 1]$$

where N is the number of documents and n is the number containing the keywords.

In this research, we propose an idea using gene description from OMIM and GeneReviews as documents and keywords as query to implement the Okapi BM25 ranking function. Therefore, we can use the Okapi BM25 score to represent the relationship between gene description and keywords. The higher score that gene description gets from keywords indicates stronger connection between that gene and keywords. Rank values were based on the Okapi BM25 ranking function mentioned before with some other parameters. Compared with the Okapi BM25 regular formula, rank value replaces the IDF function with Robertson-Spärck-Jones weight [36]. The IDF function is a measure of how much information the word provides, that is, whether the word is common or rare across all documents. For example, the term “the” is very common in every document, so term frequency will be inclined to falsely highlight the documents that happen to use the word “the” more frequently. Hence, the IDF function is dedicated to reducing the weight of words that appear very frequently among all documents. In contrast to the regular IDF function, the Robertson-Spärck-Jones weight adds relevant parameters of documents and increases the precision of rank score.

We get the phenotype-gene similarity score of each SNV from Variant Prioritizer, a text mining tool that outputs the rank and score of genes by entering symptoms as keywords. Variant Prioritizer uses the Okapi BM25 ranking function [37] to construct the connections between genes and keywords. Gene descriptions from OMIM, GeneReviews, Entrez Gene [38], and PubTator [39] serve as data sources and keywords as query to implement the Okapi BM25 score using the full-text search method. It returns a column called RANK that includes ordinal value from 0 to 1000. The RANK score is based on the following formula:

$$\text{RANK score} = \sum \omega \left[\frac{(k_1 + 1)\ tf}{K + tf}\right]\left[\frac{(k_3 + 1)\ qtf}{k_3 + qtf}\right]$$

where ω is the Robertson-Spärck-Jones weight [36], which is defined as $\omega = \log\ [(r + 0.5) \cdot (N - n - R + r + 0.5)]/[(R - r + 0.5) \cdot (n - r + 0.5)]$, in which R is the number of known relevant documents and r is the number of these containing the term; tf is the frequency of the word in the property queried within an article; qtf is the frequency of the term in the query; and K is defined as follows:

$$K = k_1[(1 - b) + b(dl/\text{avgdl})]$$

where dl is the property length, in word occurrence; avgdl is the average length of the property being queried, in word occurrence; and k_1, b, and k_3 are constants (=1.2, 0.75, and 8.0, respectively).

We employed the Variant Prioritizer API to get the RANK value from each data source as gene similarity score to represent the association between each SNVs and extracted keywords. We kept the maximum and minimum scores of rank values (4 overall) as 2 separate features for model building.

Ethical Considerations

This retrospective cohort study was approved by the Institutional Review Board (IRB) of the National Taiwan University Hospital (IRB number: 201710066RINB). We confirm that all experiments were performed in accordance with relevant guidelines and regulations. The data retrieved from EHRs were deidentified and could not be linked to the patients’ identity by the research team. The need for written informed consent was waived and confirmed by the National Taiwan University Hospital IRB (201710066RINB) because this was a retrospective cohort study with deidentified data. This regulation complies to Health Insurance Portability and Accountability Act (HIPAA) that there are no restrictions on the use or disclosure of deidentified health information.

Data Preprocessing

Overview of Steps

After variant annotation of the VCF file, we preprocessed our data into a model-acceptable format. Data preprocessing is an extremely important step in machine learning because the quality of data can directly affect the ability of a model to learn. It includes various operations and each operation aims to help machine learning build better predictive models. The data preprocessing operations used in this research are explained in the following sections.

Missing Value Handling

In real world, the data usually have missing values. AsFor example, in the genotype variable most machine learning methods cannot deal with null value, it is pivotal to identify and correctly handle the missing values. Basically, the missing values can be handled using various techniques such as deletion or imputation [40]. Deletion removes all data for an observation that has 1 or more missing values. However, if there are many columns with missing values, then deletion will result in the lack of data. Therefore, for some columns we used imputation by substituting the missing values in our data set with mean and for some columns we just simply replaced the missing value with a valid value such as 0.

One Hot Encoding

Many machine learning algorithms cannot operate on categorical data directly. They require all input features to be numeric. Basically, categorical data contain label values rather than numeric values. As a consequence, categorical data must be converted into a numerical form so that they can be used in the machine leaning model. One hot encoding is a widespread approach for dealing with categorical data. One hot encoding transforms a categorical column to a multidimensional vector. It creates new columns, indicating the presence of each possible value from the original data.

For example, in the genotype variable, there are 3 categories: homozygous (hom), heterozygous (het), and hemizygous (hem). Therefore, 3 binary variables [hom, het, hem] are needed. If genotype of a variant is heterozygous, we use [0,1,0] to represent it.

Data Normalization

For continuous data, there are a few with different ranges. If we apply features in very different ranges to some machine learning models such as logistic regression, the feature with broader range will intrinsically influence the result more owing to its larger value. However, this does not necessarily mean that this feature is more important as a predictor. Therefore, we used normalization techniques as a solution to overcome this problem. Normalization is the rescaling of the data from the original range so that all values are within the range of 0 and 1. We rescale all continuous values by min-max normalization. The general formula is as follows:

$$Xnorm = (X - Xmin)/(Xmax - Xmin)$$

where X is the original value and Xnorm is the normalized value. This will make the maximal value map to 1 and the minimal value map to 0. In addition to the aforesaid data preprocessing techniques, we handled different data types in different ways and created some new features for model building. In the following sections, we elaborate on each data type preprocessing and combine them in the end.

Functional Prediction Score

Functional prediction scores including SIFT [17], PolyPhen2 HDIV [16], PolyPhen2 HVAR [16], LRT [41], MutationTaster [18], MutationAssessor [42], FATHMM [43], PROVEAN [44], MetaSVM [14], MetaLR [14], M-CAP [45], CADD [13], GERP++ [46], DANN [47], fathmm-MKL [48], GenoCanyon [49], fitCons [50], PhyloP [51], PhastCons [52], and SiPhy [53] were from ANNOVAR. We used converted rank scores provided by ANNOVAR instead of the original score because all these scores are always within the range of 0 and 1. Besides, converted rank scores from different algorithms are monotonic in the same direction. That is, a higher score indicates that the variant is more likely to be damaging [54]. For splice site prediction, we imported the MaxEntScan score using the VEP plugin. We defined a new column called MaxEntScan significance. The value is 1 when the value of MaxEntScan alt is smaller than 3 and MaxEntScan variation is smaller than 30%; otherwise the value is 0. We used clinical significance reported in ClinVar and computed rank score from the HGMD. The HGMD computed rank score is a probability of pathogenicity between 0 and 1, with 1 being most likely disease causing compared with other HGMD entries.

Clinical Interpretation

We employed clinical interpretation of each genetic variant based on the American College of Medical Genetics and Genomics/Association for Molecular Pathology (ACMG/AMP) 2015 guideline, which is generated by InterVar. We calculated the ACMG score developed by Xrare to represent its overall pathogenicity. The ACMG score is a weighted sum score based on multiple evidence (n=14) with the following weights for each term: PVS1:6, PS1:4, PM1:2, PM2:2, PM4:2, PM5:2, PP2:1, PP3:1, BA1:9, BS1:3, BS2:3, BP3:1, BP4:1, BP7:2 [9]. We collected gene-level constraint features including pLI, pRec, syn_z, and mis_z from the Genome Aggregation Database (gnomAD). We used the patterns of inheritance defined in OMIM as our data. For variants that contain multiple patterns, we calculated the occurrences of each pattern and stored it as a feature. We also get some additional information about each variant from ANNOVAR such as genotype, regions that a variant hits, and read depths. The quality of each variant is also collected from the VCF file. As the genotype annotated by ANNOVAR does not contain hemizygous alleles, we replaced the genotype feature of all male patients' chromosome X with hemizygous alleles. In addition, we collected functional consequence on the different transcripts for a gene or in proximal regulatory regions using Nirvana.

Labels

The goal of our research was to identify the disease-causing variants with SNVs (ie, we classify a variant as disease causing or not). As machine learning algorithms learn how to assign a class label to a test case from examples, it is necessary to assign a class label to all input training sets. We used the 0/1 label to represent whether a variant is disease causing or not. If a variant is causative, we assigned label 1 to it; otherwise the label is 0. Details about all the features used in our model are presented in Multimedia Appendix 2.

Feature Selection

After data preprocessing, we got 94 features for each variant. To reduce the high dimension of the input data set while retaining the discriminatory information for classification problems, we applied univariate feature selection techniques from scikit-learn [55] packages to identify the relevant variables in a data set and eliminate the useless variables. Feature selection helps to reduce the noise in the data set and lets the model focus on the relevant signals.

There are several scoring functions provided by scikit-learn univariate feature selection modules. We used mutual information classifier to select the most relevant variables. Mutual information [56] between 2 random variables is a nonnegative value, which measures the general dependence of variables without making any assumptions about the nature of their underlying relationships [57]. The mutual information between 2 discrete random variables X and Y is defined as follows:

$$I(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log [p(x, y)/p(x) \times p(y)]$$

where $p(x, y)$ is the joint probability density function of X and Y, and $p(x)$ and $p(y)$ are the marginal density function. The mutual information determines the similarity between the joint distribution $p(x, y)$ and the products of the factored marginal distributions. The larger the value means the greater the relationship between the 2 variables. The calculated value is equal to 0 if and only if the 2 variables are independent.

We performed the feature selection process using only the training set to determine the relevant variable. Further, the number of features we selected is based on model evaluation with 10fold cross validation

Building Model

To construct a model by machine learning algorithm, we split the data into 2 groups. As our model aims to assist physicians with their clinical exome data interpretation process, the exome data from the dbGaP database and the targeted gene panel sequencing data from NTUH were set as training set, and the WES data from NTUH were set as testing data. which can only be used on model evaluation. The external validation set consisted of 90 most recent NTUH WES data, which help to make sure that our model can make predictions in future clinical use. Details about the training and testing sets are listed in Table 2.

To build the machine learning model, we implemented the random forests algorithm [58] provided by scikitlearn packages. The selection of hyperparameters is based on a grid search with 10fold cross validation. Random forest was first proposed by Leo Breiman in 2001 [58]. It is an ensemble classifier that evolves from decision trees. Actually, random forests are a combination of decision trees such that each tree depends on the values of a random vector sampled independently, with the same distribution for all trees in the forest [59]. A forest of trees is grown as follows:

- The training set is a bootstrap sample from the original training set.
- The number of trees to build and the number of variables randomly sampled as candidates at each split m-try are set by the user, where m-try is less than the total number of variables.
- At each node, m-try variables are selected at random, and the node is split on the best split point among m-try. This process iterates until the tree grows to its maximal depth.
- For test case prediction, as a test vector **x** is put down at each tree, it is assigned the average of **y** values at the node it stops at. The average of these overall trees in the forest is the predicted value for **x**. The predicted value for classification is the class getting the plurality of the forest votes..

The function we used to measure the quality of a split is Gini impurity. Gini impurity is the probability of incorrectly classifying a randomly chosen element in the data set if it were randomly labeled according to the class distribution in the data set [60]. In decision tree learning it is defined as $\mathrm{IG}(t) = \sum_{i=1}^{c} p(i|t) \times [1 - p(i|t)]$, where c is the number of classes and $p(i|t)$ is the probability of randomly picking an object of class i at node t. The optimal split from a root node when training a decision tree is chosen by maximizing the Gini gain, which is calculated by subtracting the weighted impurities of the branches from the original impurity.

Table 2. The training, testing, and external validation sets used in this study.

Data	Training set	Testing set	External validation set
Source	dbGaP[a], NTUH[b] panel	NTUH WES[c]	New NTUH WES
Patients, n	381	108	90
Filtered variants, n	125,693	80,083	109,857
Causative variants, n	478	134	100

[a] dbGaP: Database of Genotypes and Phenotypes.

[b] NTUH: National Taiwan University Hospital.

[c] WES: whole-exome sequencing.

Performance Evaluation

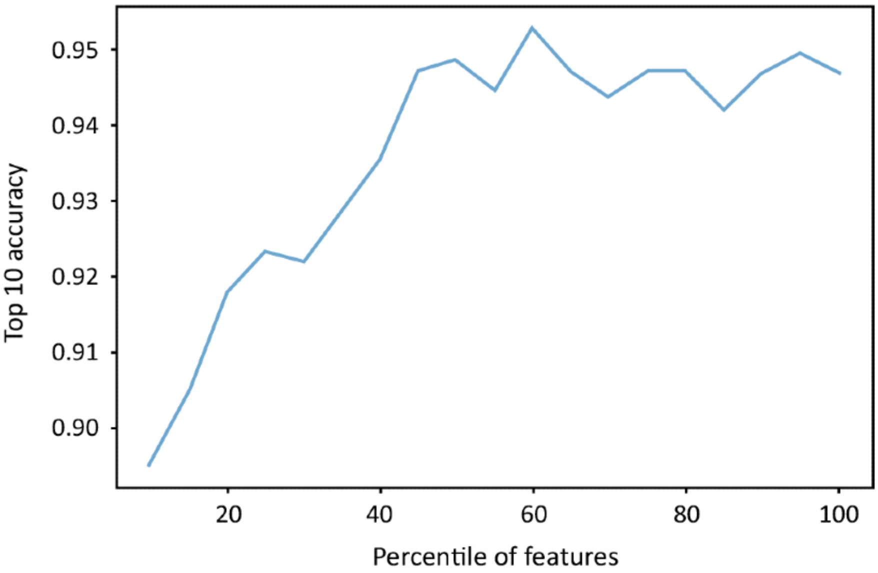
To evaluate our model performance of true causative variant prioritization, we used the ranking statistics mentioned in VarSight. After we applied the binary classification process to all variants, we got a probability for each variant that represents the probability of this variant to be disease causing. We ranked the variants for each patient from the highest to lowest probability and quantified the percentage of the target variants that were ranked in the top 1, 5, 10, 20.

Results

Feature Selection

For the feature selection, we used univariate feature selection based on the SelectPercentile method in scikitlearn package. The classifier we chose is the mutual information classifier. Only the training set was used for selecting the most relevant features. Further, we applied 10fold cross validation to decide the number of features for model training. In Figure 2, we present the top 10 accuracy on 10fold cross validation using different percentages of features. As using 60% of features achieves the highest accuracy, 56 features (60% of total 94 features) with the highest estimated mutual information were selected for the final model building.

Figure 2. The top 10 accuracy on 10-fold cross validation using different percentage of features.



Model Performance

We evaluated the model with our testing set. As mentioned in Table 2, the testing set comprised 108 patients who received WES with at least one disease-causing variant diagnosed by doctors. Multimedia Appendix 3 presents detailed information about their causative variants, keywords, and the corresponding HPO term. The keywords and HPO term are determined by doctors based on the phenotype of each patient.

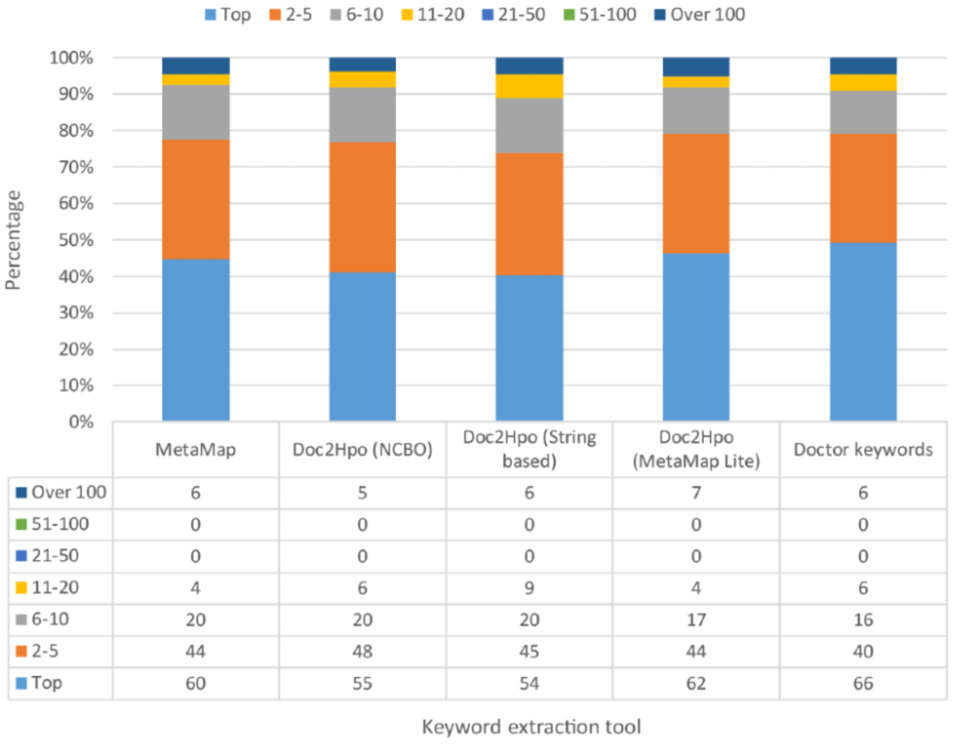
Prediction With Different Keyword Extraction Tools

Figure 3 shows the percentage distribution of the ranking of target variants and Figure 4 shows the cumulative rank result of models using different keyword extraction tools. When using tools to extract phenotypes from abstracts, our model can assign the target variants to the top rank for over 40% (60/134, 44.8%) of the total variants. The top 10 accuracies of models are around 90% (124/134, 92.5%), irrespective of the keyword extraction tool used. Compared with the keywords provided by professional doctors, applying tools to extract keywords had lower top 1 accuracy but comparable top 10 accuracy. This indicated that in most cases our model can successfully rank the true causative variants in the front of the variant lists and the rank is slightly influenced by the input keywords.

We built a random forest model based on the method described in the previous section and evaluated it with our testing set based on different keyword extraction tools. We succeeded in locating 92.5% (124/134) of the causative variant in the top 10 ranking list among an average of 741 candidate variants per person after filtering. The performance of the model is similar to that of manual analysis, and it has been used to help National Taiwan University Hospital with a genetic diagnosis.

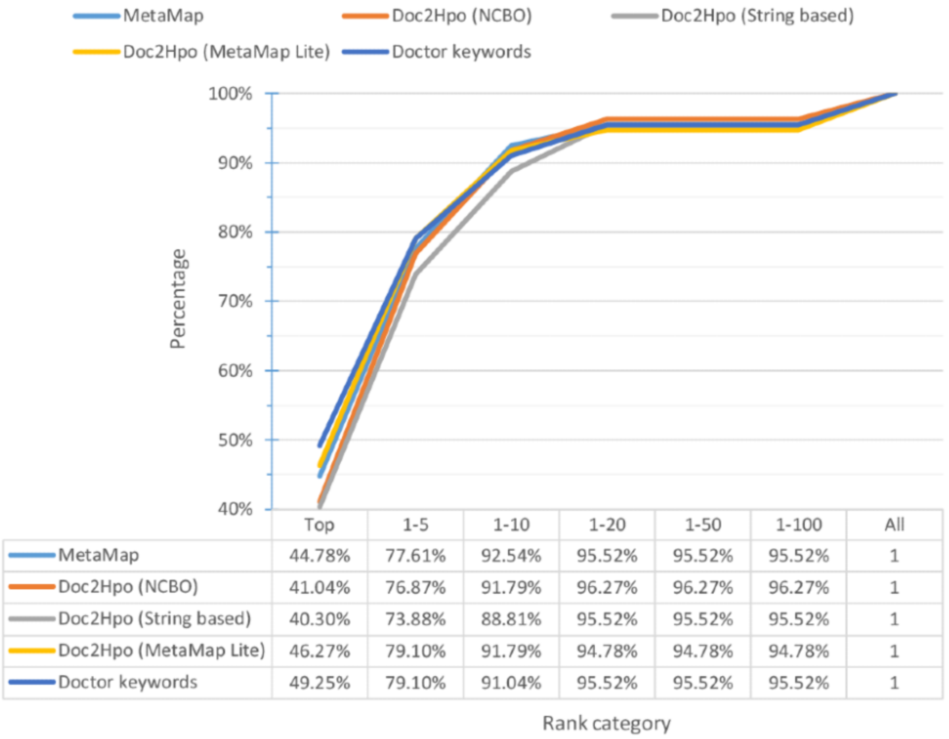
Figures 3 and 4 show the percentage distribution of the ranking of target variants and the cumulative rank result of models using different keyword extraction tools, respectively. When using tools to extract phenotypes from abstracts, our model can assign the target variants to the top rank for over 40% (60/134, 44.8%) of the total variants. The top 10 accuracies of models are around 90% (124/134, 92.5%), irrespective of the keyword extraction tool used. Compared with the keywords provided by professional doctors, applying tools to extract keywords has lower top 1 accuracy but comparable top 10 accuracy. It represents that in most cases our model can successfully rank the true causative variants in the front of the variant lists and the rank is slightly influenced by the input keywords.

Figure 3. Percentage distribution of ranks.



	MetaMap	Doc2Hpo (NCBO)	Doc2Hpo (String based)	Doc2Hpo (MetaMap Lite)	Doctor keywords
Over 100	6	5	6	7	6
51-100	0	0	0	0	0
21-50	0	0	0	0	0
11-20	4	6	9	4	6
6-10	20	20	20	17	16
2-5	44	48	45	44	40
Top	60	55	54	62	66

Figure 4. Cumulative percentage distribution of ranks. NCBO: National Center for Biomedical Ontology.

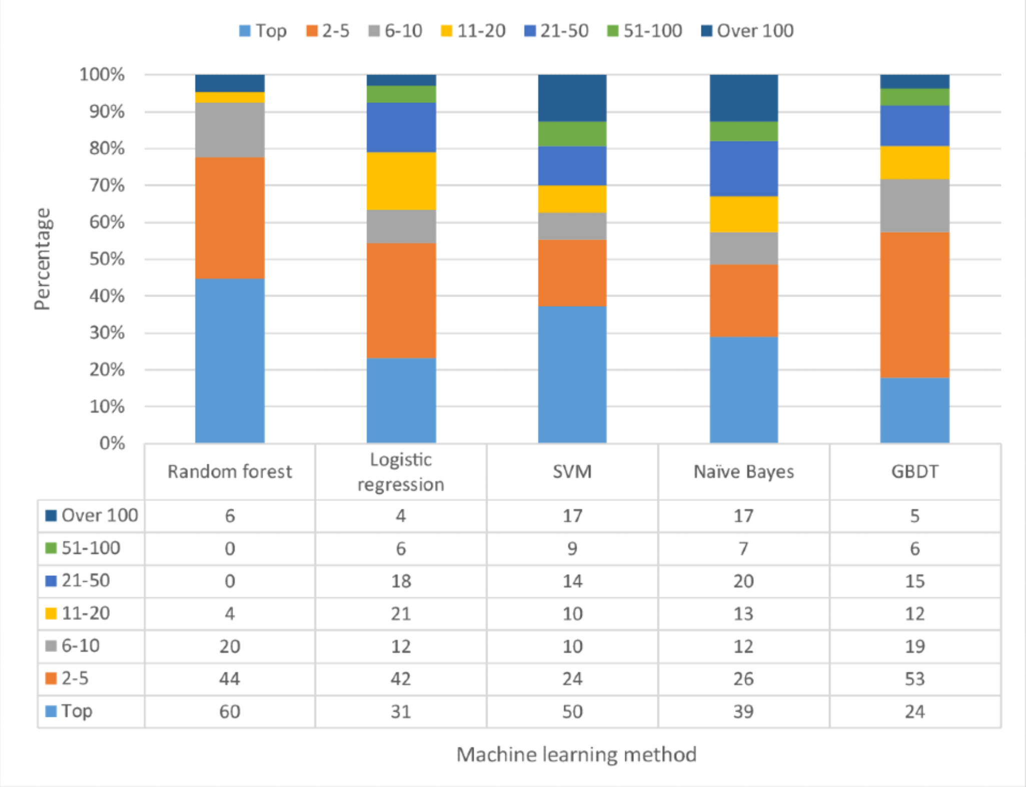


	Top	1-5	1-10	1-20	1-50	1-100	All
MetaMap	44.78%	77.61%	92.54%	95.52%	95.52%	95.52%	1
Doc2Hpo (NCBO)	41.04%	76.87%	91.79%	96.27%	96.27%	96.27%	1
Doc2Hpo (String based)	40.30%	73.88%	88.81%	95.52%	95.52%	95.52%	1
Doc2Hpo (MetaMap Lite)	46.27%	79.10%	91.79%	94.78%	94.78%	94.78%	1
Doctor keywords	49.25%	79.10%	91.04%	95.52%	95.52%	95.52%	1

Other Machine Learning Methods

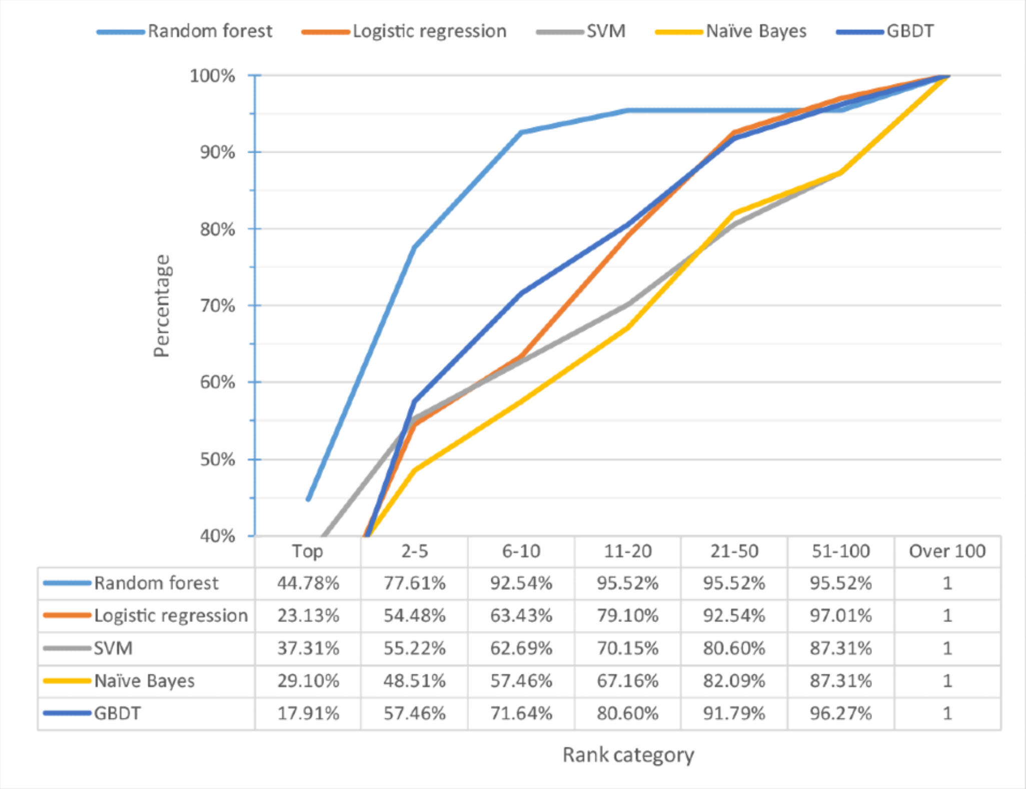
We also evaluated other machine learning methods and compared their performance with random forest. These methods include logistic regression, Gaussian naive Bayes, SVM with RBF kernel, and gradient boosted decision trees. The selection of hyperparameters for each algorithm was based on grid search with 10-fold cross validation. We used MetaMap as the keyword extraction tool and our testing data to test the performance of each method. The percentage distribution of the ranking of target variants by each machine learning method and the cumulative rank result of each model are shown in Figures 5 and 6, respectively. As random forest got the highest top 10 accuracy, we finally chose random forest as our machine learning algorithm.

Figure 5. Percentage distribution of ranks. GBDT: gradient boosting decision tree; SVM: support vector machine.



	Random forest	Logistic regression	SVM	Naïve Bayes	GBDT
■ Over 100	6	4	17	17	5
■ 51-100	0	6	9	7	6
■ 21-50	0	18	14	20	15
■ 11-20	4	21	10	13	12
■ 6-10	20	12	10	12	19
■ 2-5	44	42	24	26	53
■ Top	60	31	50	39	24

Figure 6. Cumulative percentage distribution of ranks. GBDT: gradient boosting decision tree; SVM: support vector machine.



	Top	2-5	6-10	11-20	21-50	51-100	Over 100
— Random forest	44.78%	77.61%	92.54%	95.52%	95.52%	95.52%	1
— Logistic regression	23.13%	54.48%	63.43%	79.10%	92.54%	97.01%	1
— SVM	37.31%	55.22%	62.69%	70.15%	80.60%	87.31%	1
— Naïve Bayes	29.10%	48.51%	57.46%	67.16%	82.09%	87.31%	1
— GBDT	17.91%	57.46%	71.64%	80.60%	91.79%	96.27%	1

Discussion

Principal Findings

We have implemented a website, AI Variant Prioritizer, which uses data from NGS bioinformatics pipelines with machine learning to make a prediction about most possible disease-causing variants among SNVs and patient’s phenotype data. This system can assist researchers and physicians by focusing on those with higher disease-causing probability and reducing the average turnaround time (by 1 day) of the entire WES pipeline, from DNA extraction to clinical diagnosis.

Moreover, we have implemented a web API for our system so that the ranking function could be integrated into MViewer. Thus, physicians can interpret the results of genetic variation with a single application instead of adopting numerous services.

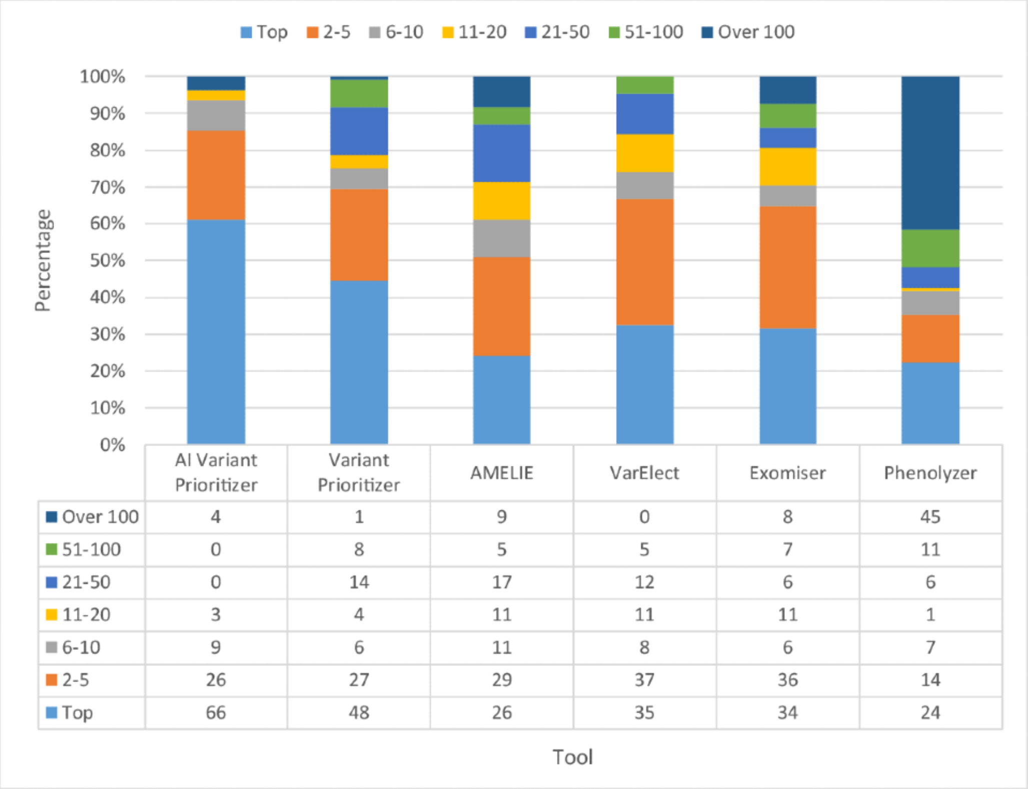
For comparison, we used our testing data to run several prioritization tools including AMELIE [61], VarElect [62], Exomiser, Phenolyzer, and Variant Prioritizer. As AMELIE and Exomiser can only accept phenotypes defined in HPO terms, we entered HPO terms determined by doctors as their input. Phenolyzer can identify both disease terms and HPO terms. However, if the terms do not match in their databases, it will not return any record. Hence, we also used HPO terms as input for Phenolyzer. VarElect, Variant Prioritizer, and our model can identify free-text descriptions. Therefore, we imputed the keywords provided by doctors as input for testing. AMELIE, VarElect, and Variant Prioritizer only prioritize the gene list instead of the variant list. Hence, we evaluated the result for gene-based prioritization instead of variant-based prioritization. That is, for each patient, if the tools prioritize target gene in the top 1, 5, 10, 20, 50, and 100 of our filtered gene lists, this patient will be counted. All the tools use the default settings provided in their websites to run.

Figures 7 and 8 show the percentage and cumulative percentage distribution of the target gene ranking for each tool, respectively. From Figure 8, we can see that AI Variant Prioritizer is able to assign the target gene to the top rank for 61.1% (66/108) of the patients, followed by Variant Prioritizer (48/108, 44.4%). It also shows the cumulative rank result, which reveals that our AI Variant Prioritizer has the highest accuracy at ranks 1, 5, 10, and 20. Further, AI Variant Prioritizer shows better performance than other tools. After adopting the HPO terms by looking up the databases, the top 10 ranking list can be increased to 93.5% (101/108).

In comparison with extraction of phenotypic features from SNOMED through manual mapping of HPO terms to SNOMED Clinical Terms (SNOMED CT) [63], our automation approach explores various phenotypic feature extraction tools and focuses on rare disease interpretation. We have also looked into several AI-driven variant prioritization approaches published in the last 3 years, including Fabric GEM [12], MOON [2], and Exomiser. There are several differences between our approach and each of these approaches, including the algorithms used to build the prioritization model, the features considered, and databases integrated. However, the major difference of our approach from others is the method used to turn the relationships between genes and phenotypes into numerical values, which makes way for later prediction. Fabric GEM and MOON utilize Phevor [15] to turn phenotype-gene relationship into numerical values, whereas Exomiser uses PhenoDigm [64] to achieve this goal.

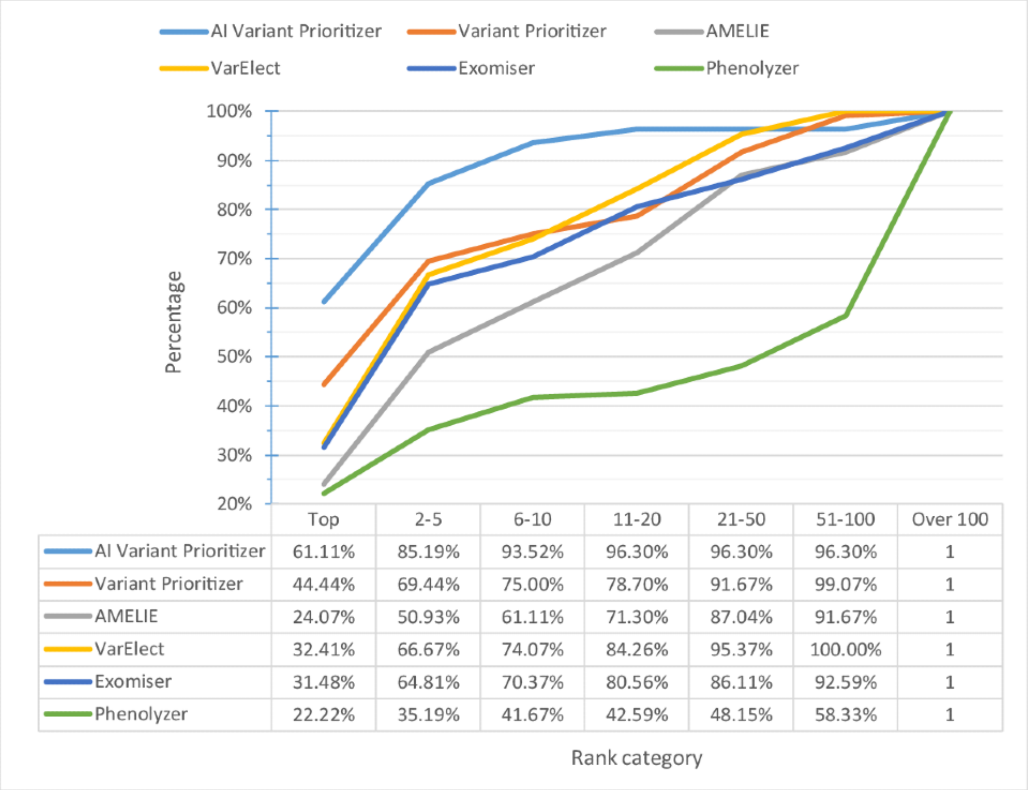
Both Phevor and PhenoDigm construct new methods that bridge HPO and other ontologies to discover more gene-disease associations. Phevor gathers all correlation of diseases and genes provided by HPO and Gene Ontology (GO) and constructs several networks (graphs) and distributes decreasing weights along the paths found. The sum of weights on the specific gene node represents the correlation score of the gene with the corresponding HPO term. PhenoDigm utilizes OWLSim [65] to calculate the similarity among different phenotypes in different ontologies and uses similarity to estimate the magnitude of correlation of given HPO terms and different genes. By contrast, Variant Prioritizer used in our approach extracts the phenotype-gene relationship from a different kind of knowledge source: free text of database. We make a simple comparison of these approaches in Tables 3 and 4.

Figure 7. Percentage distribution of ranks. AI: artificial intelligence.



	AI Variant Prioritizer	Variant Prioritizer	AMELIE	VarElect	Exomiser	Phenolyzer
■ Over 100	4	1	9	0	8	45
■ 51-100	0	8	5	5	7	11
■ 21-50	0	14	17	12	6	6
■ 11-20	3	4	11	11	11	1
■ 6-10	9	6	11	8	6	7
■ 2-5	26	27	29	37	36	14
■ Top	66	48	26	35	34	24

Figure 8. Cumulative percentage distribution of ranks. AI: artificial intelligence.



	Top	2-5	6-10	11-20	21-50	51-100	Over 100
AI Variant Prioritizer	61.11%	85.19%	93.52%	96.30%	96.30%	96.30%	1
Variant Prioritizer	44.44%	69.44%	75.00%	78.70%	91.67%	99.07%	1
AMELIE	24.07%	50.93%	61.11%	71.30%	87.04%	91.67%	1
VarElect	32.41%	66.67%	74.07%	84.26%	95.37%	100.00%	1
Exomiser	31.48%	64.81%	70.37%	80.56%	86.11%	92.59%	1
Phenolyzer	22.22%	35.19%	41.67%	42.59%	48.15%	58.33%	1

Table 3. The comparison among AI Variant Prioritizer, Fabric GEM, MOON, and Exomiser.

Tool	AI[a] Variant Prioritizer	Fabric GEM	MOON	Exomiser
Variant scoring algorithm	Random forest	Bayes factor	Decision trees, Bayesian models, neural networks	Rule based
Phenotype-gene score	Variant Prioritizer	Phevor	Phevor	PhenoDigm
Phenotype input format	Plain texts	HPO[b] terms	HPO terms extracted from electronic health record	HPO terms

[a] AI: artificial intelligence.

[b] HPO: Human Phenotype Ontology.

Table 4. The comparison among Variant Prioritizer, Phevor, and PhenoDigm.

Tool	Variant Prioritizer	Phevor	PhenoDigm
Algorithm	Okapi BM25	Graph algorithm	OWLSim
Phenotype input format	Plain texts	HPO[a] terms	HPO terms
Knowledge base	OMIM[b], GeneReviews, Entrez Gene and PubTator	HPO and GO[c]	OMIM (HPO), Sanger-MGP [66], MGD [67], and ZFIN [68]

[a] HPO: Human Phenotype Ontology.

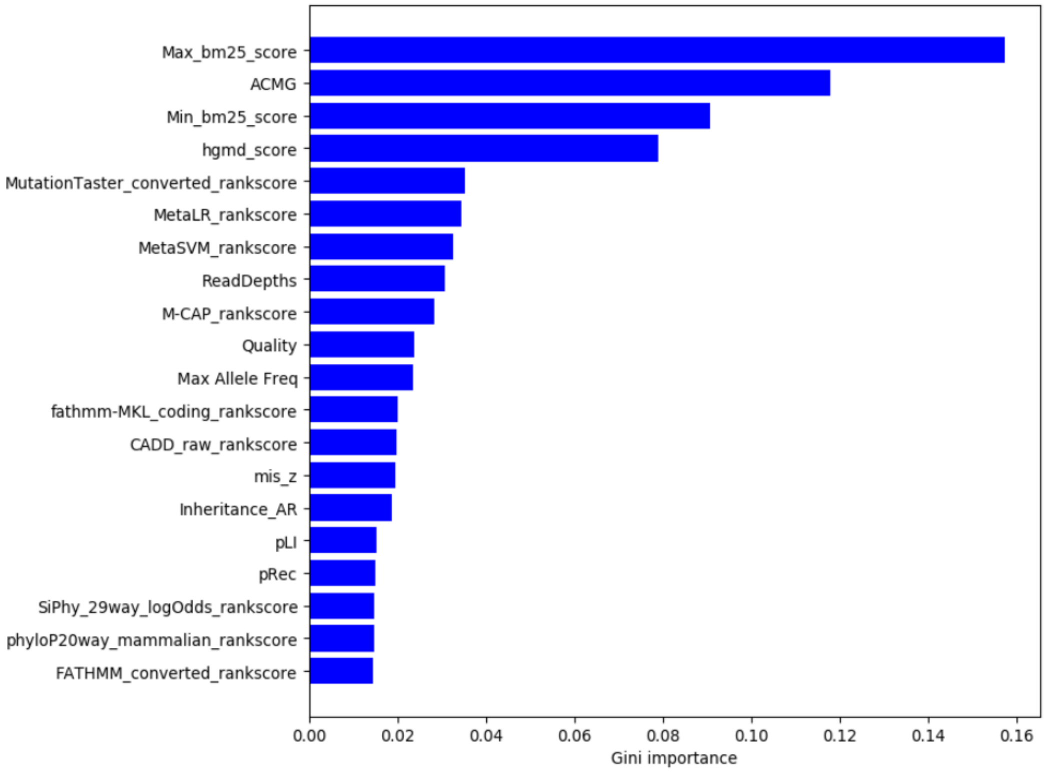
[b] OMIM: Online Mendelian Inheritance in Man.

[c] GO: Gene Ontology.

Feature Importance

For interpreting model predictions, we used the feature importance method provided by scikit-learn to identify which feature has the most predictive power. Figure 9 shows the top 20 important features. According to clinical experience, the connection between a variant and phenotype of a patient is a key factor that influences the physician to decide whether to report a variant or not. Similarly, from the figure we can see that the most important feature is the max bm25 score, which refers to the similarity score between the given variant and keywords. Another important factor that influences the reporting decision during clinical analysis is the severity of a variant. The corresponding feature we use is the ACMG score, which is in the second place of feature importance. By contrast, the result of feature importance might provide information for physicians or researchers about the features that they can consider when finding causative variant.

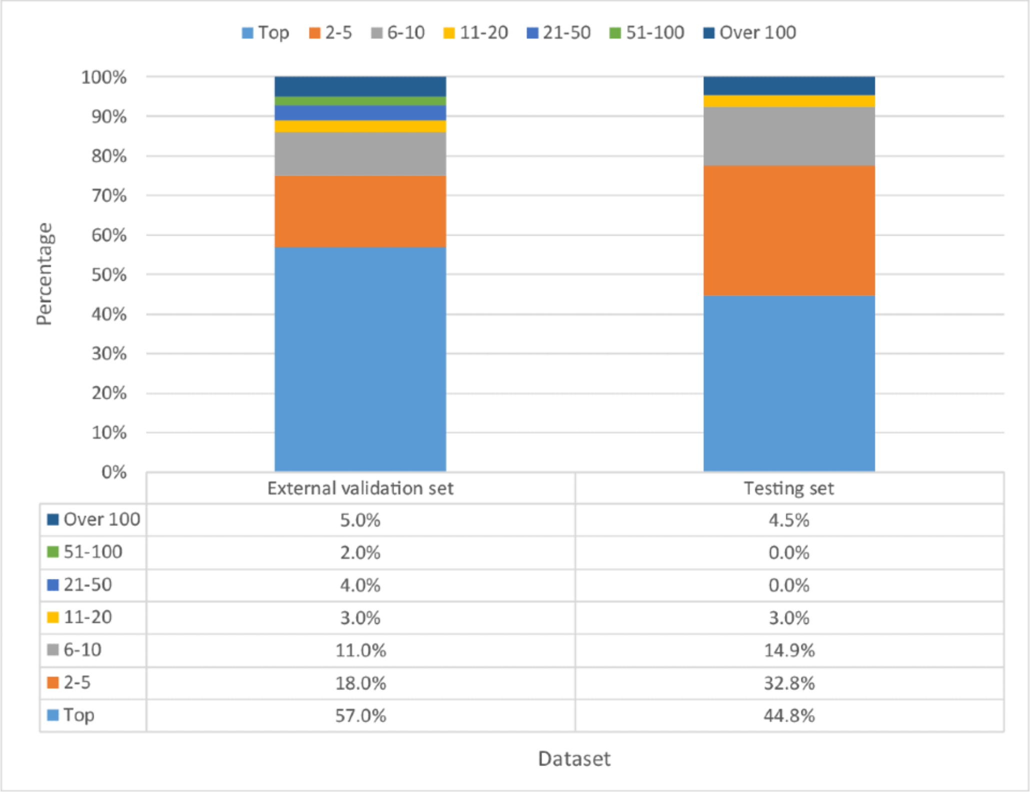
Figure 9. Feature importance.



External Validation

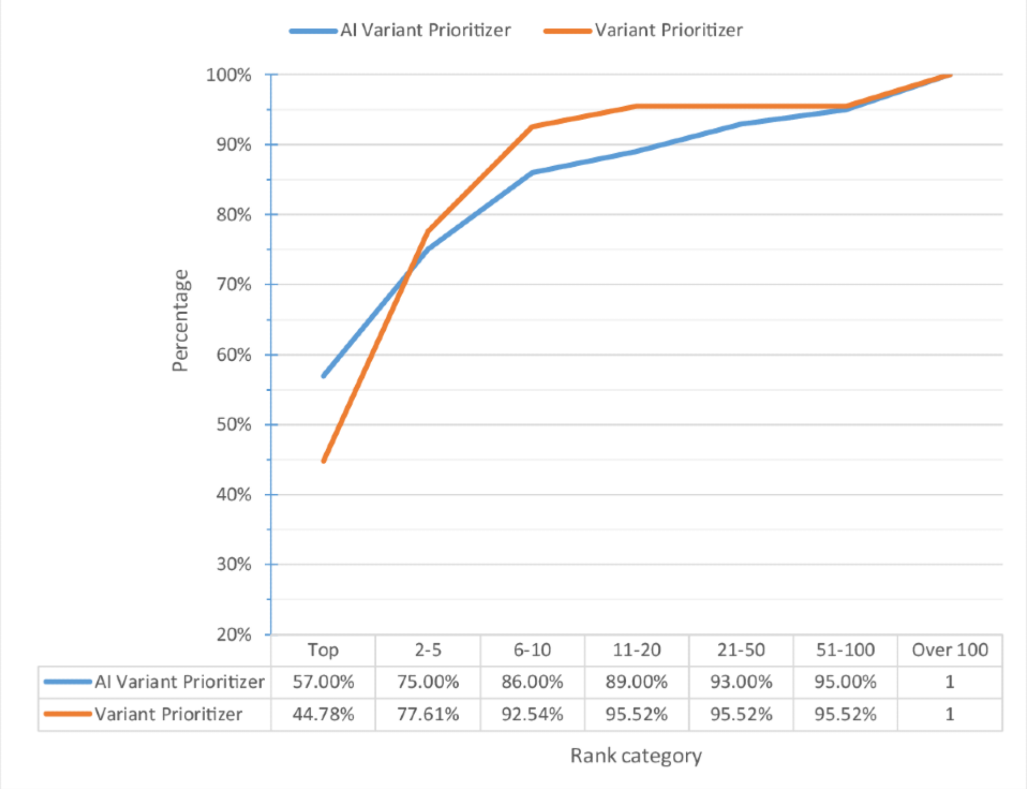
We compared the cumulative percentage distribution of ranks from the testing set and the external validation set. The result is shown in Figures 10 and 11. Their percentage values are close to each other in different regions such as top 10 and top 5. The percentage of top 1 rank of the external validation set is even higher than that of the testing set. With this result, we believe that our approach has shown its potential for robust clinical usage.

Figure 10. Percentage distribution of ranks.



	External validation set	Testing set
■ Over 100	5.0%	4.5%
■ 51-100	2.0%	0.0%
■ 21-50	4.0%	0.0%
■ 11-20	3.0%	3.0%
■ 6-10	11.0%	14.9%
■ 2-5	18.0%	32.8%
■ Top	57.0%	44.8%

Figure 11. Cumulative percentage distribution of ranks.



	Top	2-5	6-10	11-20	21-50	51-100	Over 100
AI Variant Prioritizer	57.00%	75.00%	86.00%	89.00%	93.00%	95.00%	1
Variant Prioritizer	44.78%	77.61%	92.54%	95.52%	95.52%	95.52%	1

Limitations

The study has several potential limitations. First, we could not find massive data for training and testing. This not only restricts the amount of teaching material for the machine learning model, but also restricts available measurements to evaluate the trained model. Second, the gene-phenotype score used in this study did not have enough power to detect small or moderate associations because it relies on how frequently the gene-phenotype relationship is reported to the databases it utilizes. Finally, the study did not adjust for potential confounders, such as diet and physical activity. This could cause potential bias because the way in which genes are expressed can be impacted by lifestyle of patients.

Overall, this study could have potential bias resulting from the lack of sufficient data, lack of reported gene-phenotype relationship, and lack of observation of lifestyle. The impact from the first and the second can be reduced if there are more data and reports available in the future. On the other side, the influence of lifestyle and environment remains an issue that needs more dedicated studies.

Conclusions

In this research, we proposed a machine learning model, AI Variant Prioritizer, to predict whether a variant is disease causing for patients with rare Mendelian disorder. We have successfully applied sequencing data from WES and free-text phenotypic information of patient's disease automatically extracted by keyword extraction tools for model training and testing. By interpreting our model, we identified which features of variants are important. Besides, we achieved a satisfactory result on finding the target variant in our testing data set. After testing 108 patients' WES data, we succeeded in 93.5% (n=101) of the cases to locate the causative variant in the top 10 ranking list among an average of 741 candidate variants per person after the filtering process. The performance of the model is similar to that of manual analysis by the physicians in the Department of Medical Genetics, NTUH, and it has been used to help NTUH with a genetic diagnosis.

As the physicians are very busy almost all the time in taking care of their patients, the search time spent for an accurate genetic diagnosis is extremely important. Our AI predicting model can provide the top 10 hit list with a high probability of 93.5% (101/108), thus helping them save weeks of time if they have to do it manually to search beyond the top 10 list very often.

It is not an easy work to fully interpret the causative variations of a genetic disease. As the precision of the keywords extracted by tools influence the performance of our model, for the future work, we will adopt some NLP techniques such as Bidirectional Encoder Representations from Transformers (BERT) to extract keywords more properly. In addition, the AI Variant Prioritizer model has been built to analyze SNVs and small indels from WES data, but we have not dealt with copy number variations (CNVs) yet. CNVs have been recognized as critical genetic variations, which are associated with both common and complex diseases, and thus have a large influence on several Mendelian and somatic genetic disorders. Therefore, we will collect data on CNVs and extend the capability of our system to annotate and filter CNVs. Furthermore, we will enlarge our data set by adding CNVs as our training data to enable the AI Variant Prioritizer model to predict any kind of causative genetic variations. Before implementation of AI Variant Prioritizer, the mean turnaround time of the entire WES pipeline, from DNA extraction to clinical diagnosis, was 5.8 (SD 1.1) days using Variant Prioritizer. However, after implementation of AI Variant Prioritizer, the mean turnaround time was reduced to 4.8 (SD 1.2) days for rapid trio exome sequencing analysis in NTUH.

Acknowledgments

This work was funded by a grant from the Ministry of Science and Technology (110-2634-F-002-032-) of Taiwan. The analyses presented in this publication are based on the use of study data downloaded from the dbGaP website, under dbGaP accession numbers phs000744.v4.p2, phs001272.v1.p1, phs000971.v2.p1, phs000711.v6.p2, and phs001232.v3.p2. W-LH applied for the data with project name "Variant prioritization for rapid exome analysis of rare genetic disease" (project ID 20911). Data were downloaded from the FTP site of dbGaP after approval.

Authors' Contributions

Y-SH investigated the model and data feasibility, performed formal analysis, developed the software, visualized data, and wrote the initial manuscript. CH conceived the idea, curated the data, reviewed the manuscript, and advised the software development team. N-CL and W-LH conceived the idea, curated the patient data, and reviewed and edited the draft. Y-CC and I-CL edited, revised, and strengthened the manuscript. HW and Y-LL tested the data performance. FPL supervised the project progress and supported the project and managed the project and reviewed the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Allele frequency.
[DOCX File , 16 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Description of features used in this research.
[XLSX File (Microsoft Excel File), 13 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Target variants, HPO term, and keywords of test case. HPO: Human Phenotype Ontology.
[XLSX File (Microsoft Excel File), 24 KB-Multimedia Appendix 3]

References

1. Behjati S, Tarpey PS. What is next generation sequencing? Arch Dis Child Educ Pract Ed. Dec 2013;98(6):236-238. [FREE Full text] [doi: 10.1136/archdischild-2013-304340] [Medline: 23986538]
2. O'Brien TD, Campbell NE, Potter AB, Letaw JH, Kulkarni A, Richards CS. Artificial intelligence (AI)-assisted exome reanalysis greatly aids in the identification of new positive cases and reduces analysis time in a clinical diagnostic laboratory. Genet Med. Jan 2022;24(1):192-200. [doi: 10.1016/j.gim.2021.09.007] [Medline: 34906498]
3. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. Dec 1977;74(12):5463-5467. [FREE Full text] [doi: 10.1073/pnas.74.12.5463] [Medline: 271968]
4. Hamosh A, Scott AF, Amberger J, Valle D, McKusick VA. Online Mendelian Inheritance in Man (OMIM). Hum Mutat. 2000;15(1):57-61. [doi: 10.1002/(SICI)1098-1004(200001)15:1<57::AID-HUMU12>3.0.CO;2-G] [Medline: 10612823]
5. Adam MP, Everman DB, Mirzaa GM, Pagon RA, Wallace SE, Bean LJH, et al, editors. GeneReviews. Seattle, WA. University of Washington, Seattle; 1993.
6. Faintuch J, Faintuch S. Precision Medicine for Investigators, Practitioners and Providers. New York, NY. Academic Press; Nov 16, 2019.
7. Smedley D, Jacobsen JOB, Jäger M, Köhler S, Holtgrewe M, Schubach M, et al. Next-generation diagnostics and disease-gene discovery with the Exomiser. Nat Protoc. Dec 2015;10(12):2004-2015. [FREE Full text] [doi: 10.1038/nprot.2015.124] [Medline: 26562621]
8. Boudellioua I, Kulmanov M, Schofield PN, Gkoutos GV, Hoehndorf R. DeepPVP: phenotype-based prioritization of causative variants using deep learning. BMC Bioinformatics. Feb 06, 2019;20(1):65. [FREE Full text] [doi: 10.1186/s12859-019-2633-8] [Medline: 30727941]
9. Li Q, Zhao K, Bustamante CD, Ma X, Wong WH. Xrare: a machine learning method jointly modeling phenotypes and genetic evidence for rare disease diagnosis. Genet Med. Sep 2019;21(9):2126-2134. [FREE Full text] [doi: 10.1038/s41436-019-0439-8] [Medline: 30675030]

10. Holt JM, Wilk B, Birch CL, Brown DM, Gajapathy M, Moss AC, Undiagnosed Diseases Network, et al. VarSight: prioritizing clinically reported variants with binary classification algorithms. BMC Bioinformatics. Oct 15, 2019;20(1):496. [FREE Full text] [doi: 10.1186/s12859-019-3026-8] [Medline: 31615419]
11. Yang H, Robinson PN, Wang K. Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. Nat Methods. Sep 2015;12(9):841-843. [FREE Full text] [doi: 10.1038/nmeth.3484] [Medline: 26192085]
12. De La Vega FM, Chowdhury S, Moore B, Frise E, McCarthy J, Hernandez EJ, et al. Artificial intelligence enables comprehensive genome interpretation and nomination of candidate diagnoses for rare genetic diseases. Genome Med. Oct 14, 2021;13(1):153. [FREE Full text] [doi: 10.1186/s13073-021-00965-0] [Medline: 34645491]
13. Rentzsch P, Witten D, Cooper G, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. Jan 08, 2019;47(D1):D886-D894. [FREE Full text] [doi: 10.1093/nar/gky1016] [Medline: 30371827]
14. Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum Mol Genet. May 15, 2015;24(8):2125-2137. [FREE Full text] [doi: 10.1093/hmg/ddu733] [Medline: 25552646]
15. Singleton MV, Guthery SL, Voelkerding KV, Chen K, Kennedy B, Margraf RL, et al. Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families. Am J Hum Genet. Apr 03, 2014;94(4):599-610. [FREE Full text] [doi: 10.1016/j.ajhg.2014.03.010] [Medline: 24702956]
16. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. Apr 2010;7(4):248-249. [FREE Full text] [doi: 10.1038/nmeth0410-248] [Medline: 20354512]
17. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. Jul 01, 2003;31(13):3812-3814. [FREE Full text] [doi: 10.1093/nar/gkg509] [Medline: 12824425]
18. Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: mutation prediction for the deep-sequencing age. Nat Methods. Apr 2014;11(4):361-362. [doi: 10.1038/nmeth.2890] [Medline: 24681721]
19. Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. Am J Hum Genet. Nov 2008;83(5):610-615. [FREE Full text] [doi: 10.1016/j.ajhg.2008.09.017] [Medline: 18950739]
20. Hsu C. An integrated genetic variation analysis system for gene diagnostics in precision medicine (Master's thesis). NDLTD. Taipei City, Taiwan. National Taiwan University; 2018. URL: https://hdl.handle.net/11296/v9rcd8 [accessed 2022-08-31]
21. ChenT-F. Variants Prioritizer for Exome Data Based on Text-mining. NTU Thesis and Dissertations Repository. Taipei City, Taiwan. National Taiwan University; 2018. URL: https://tdr.lib.ntu.edu.tw/handle/123456789/17687?mode=full [accessed 2022-08-31]
22. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. Sep 2010;38(16):e164. [FREE Full text] [doi: 10.1093/nar/gkq603] [Medline: 20601685]
23. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. Jun 06, 2016;17(1):122. [FREE Full text] [doi: 10.1186/s13059-016-0974-4] [Medline: 27268795]
24. Stromberg M, Roy R, Lajugie J, Jiang Y, Li H, Margulies E. Nirvana: Clinical Grade Variant Annotator. New York, NY. Association for Computing Machinery; 2017. Presented at: The 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics; August 20-23, 2017; Boston, MA. [doi: 10.1145/3107411.3108204]
25. Li Q, Wang K. InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. Am J Hum Genet. Feb 02, 2017;100(2):267-280. [FREE Full text] [doi: 10.1016/j.ajhg.2017.01.004] [Medline: 28132688]
26. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. Jan 2014;42(Database issue):D980-D985. [FREE Full text] [doi: 10.1093/nar/gkt1113] [Medline: 24234437]
27. Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NST, et al. Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat. Jul 2003;21(6):577-581. [doi: 10.1002/humu.10212] [Medline: 12754702]
28. Fan C, Lin J, Lee C. Taiwan Biobank: a project aiming to aid Taiwan's transition into a biomedical island. Pharmacogenomics. Feb 2008;9(2):235-246. [doi: 10.2217/14622416.9.2.235] [Medline: 18370851]
29. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. ACMG Laboratory Quality Assurance Committee. 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. Genet Med. May 2015;17(5):405-424. [FREE Full text] [doi: 10.1038/gim.2015.30] [Medline: 25741868]
30. Glassberg EC, Gao Z, Harpak A, Lant X, Pritchard JK. Measurement of selective constraint on human gene expression. bioRxiv. 2022. [doi: 10.1101/345801]
31. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 2001:17-21. [FREE Full text] [Medline: 11825149]
32. Aronson AR, Lang F. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc. 2010;17(3):229-236. [FREE Full text] [doi: 10.1136/jamia.2009.002733] [Medline: 20442139]

33. Lindberg DAB, Humphreys BL, McCray AT. The Unified Medical Language System. Yearb Med Inform. Mar 05, 2018;02(01):41-51. [doi: 10.1055/s-0038-1637976]
34. Liu C, Peres Kury FS, Li Z, Ta C, Wang K, Weng C. Doc2Hpo: a web application for efficient and accurate HPO concept curation. Nucleic Acids Res. Jul 02, 2019;47(W1):W566-W570. [FREE Full text] [doi: 10.1093/nar/gkz386] [Medline: 31106327]
35. Tchechmedjiev A, Abdaoui A, Emonet V, Melzi S, Jonnagaddala J, Jonquet C. Enhanced functionalities for annotating and indexing clinical text with the NCBO Annotator. Bioinformatics. Jun 01, 2018;34(11):1962-1965. [FREE Full text] [doi: 10.1093/bioinformatics/bty009] [Medline: 29846492]
36. Lee L. IDF revisited: A simple new derivation within the Robertson-Spärck Jones probabilistic model. New York, NY. ACM; 2007. Presented at: SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information; July 23-27, 2007:751-752; Amsterdam, The Netherlands. [doi: 10.1145/1277741.1277891]
37. Robertson S, Walker S, Beaulieu MM. Okapi at TREC-7: automatic ad hoc, filtering, VCL and interactive track. Microsoft. Gaithersburg, MD. National Institute of Standards and Technology; Jan 1999. URL: https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/okapi_trec7.pdf [accessed 2022-08-31]
38. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res. Jan 01, 2005;33(Database issue):D54-D58. [FREE Full text] [doi: 10.1093/nar/gki031] [Medline: 15608257]
39. Wei C, Kao H, Lu Z. PubTator: a web-based text mining tool for assisting biocuration. Nucleic Acids Res. Jul 2013;41(Web Server issue):W518-W522. [FREE Full text] [doi: 10.1093/nar/gkt441] [Medline: 23703206]
40. Hintzsche JD, Robinson WA, Tan AC. A Survey of Computational Tools to Analyze and Interpret Whole Exome Sequencing Data. Int J Genomics. 2016;2016:7983236. [FREE Full text] [doi: 10.1155/2016/7983236] [Medline: 28070503]
41. Doniger SW, Kim HS, Swain D, Corcuera D, Williams M, Yang S, et al. A catalog of neutral and deleterious polymorphism in yeast. PLoS Genet. Aug 29, 2008;4(8):e1000183. [FREE Full text] [doi: 10.1371/journal.pgen.1000183] [Medline: 18769710]
42. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. Sep 01, 2011;39(17):e118. [FREE Full text] [doi: 10.1093/nar/gkr407] [Medline: 21727090]
43. Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GLA, Edwards KJ, et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Hum Mutat. Jan 2013;34(1):57-65. [FREE Full text] [doi: 10.1002/humu.22225] [Medline: 23033316]
44. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. PLoS One. 2012;7(10):e46688. [FREE Full text] [doi: 10.1371/journal.pone.0046688] [Medline: 23056405]
45. Jagadeesh KA, Wenger AM, Berger MJ, Guturu H, Stenson PD, Cooper DN, et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. Nat Genet. Oct 24, 2016;48(12):1581-1586. [doi: 10.1038/ng.3703]
46. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol. Dec 02, 2010;6(12):e1001025. [FREE Full text] [doi: 10.1371/journal.pcbi.1001025] [Medline: 21152010]
47. Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics. Mar 01, 2015;31(5):761-763. [FREE Full text] [doi: 10.1093/bioinformatics/btu703] [Medline: 25338716]
48. Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day INM, et al. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics. May 15, 2015;31(10):1536-1543. [FREE Full text] [doi: 10.1093/bioinformatics/btv009] [Medline: 25583119]
49. Lu Q, Hu Y, Sun J, Cheng Y, Cheung K, Zhao H. A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. Sci Rep. May 27, 2015;5(1):10576-10513. [FREE Full text] [doi: 10.1038/srep10576] [Medline: 26015273]
50. Gulko B, Hubisz MJ, Gronau I, Siepel A. A method for calculating probabilities of fitness consequences for point mutations across the human genome. Nat Genet. Mar 2015;47(3):276-283. [FREE Full text] [doi: 10.1038/ng.3196] [Medline: 25599402]
51. Hubisz MJ, Pollard KS, Siepel A. PHAST and RPHAST: phylogenetic analysis with space/time models. Brief Bioinform. Jan 2011;12(1):41-51. [FREE Full text] [doi: 10.1093/bib/bbq072] [Medline: 21278375]
52. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. Aug 15, 2005;15(8):1034-1050. [FREE Full text] [doi: 10.1101/gr.3715005] [Medline: 16024819]
53. Garber M, Guttman M, Clamp M, Zody M, Friedman N, Xie X. Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics. Jul 15, 2009;25(12):i54-i62. [FREE Full text] [doi: 10.1093/bioinformatics/btp190] [Medline: 19478016]
54. Liu X, Wu C, Li C, Boerwinkle E. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. Hum Mutat. Mar 2016;37(3):235-241. [FREE Full text] [doi: 10.1002/humu.22932] [Medline: 26555599]

55. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O. Scikit-learn: machine learning in python. Journal of Machine Learning Research. 2011;12:2825-2830. [FREE Full text]
56. Kraskov A, Stögbauer H, Grassberger P. Estimating mutual information. Phys. Rev. E. Jun 23, 2004;69(6):066138-1-066138-16. [doi: 10.1103/physreve.69.066138]
57. Ross BC. Mutual information between discrete and continuous data sets. PLoS One. Feb 19, 2014;9(2):e87357. [FREE Full text] [doi: 10.1371/journal.pone.0087357] [Medline: 24586270]
58. Breiman L. Random forests. Machine Learning. 2001;45:5-32. [FREE Full text] [doi: 10.1023/A:1010933404324]
59. Breiman L. Consistency for a simple model of random forests. University of California, Berkeley. 2004. URL: https://www.stat.berkeley.edu/~breiman/RandomForests/consistencyRFA.pdf [accessed 2022-08-31]
60. Ellerman D. Logical Entropy: Introduction to Classical and Quantum Logical Information Theory. Entropy (Basel). Oct 06, 2018;20(9):679. [FREE Full text] [doi: 10.3390/e20090679] [Medline: 33265768]
61. Birgmeier J, Haeussler M, Deisseroth CA, Jagadeesh KA, Ratner AJ, Guturu H, et al. AMELIE accelerates Mendelian patient diagnosis directly from the primary literature. bioRxiv. Preprint posted online on August 02, 2017. [doi: 10.1101/171322]
62. Stelzer G, Plaschkes I, Oz-Levi D, Alkelai A, Olender T, Zimmerman S, et al. VarElect: the phenotype-based variation prioritizer of the GeneCards Suite. BMC Genomics. Jun 23, 2016;17 Suppl 2(S2):444-206. [FREE Full text] [doi: 10.1186/s12864-016-2722-2] [Medline: 27357693]
63. Clark MM, Hildreth A, Batalov S, Ding Y, Chowdhury S, Watkins K, et al. Diagnosis of genetic diseases in seriously ill children by rapid whole-genome sequencing and automated phenotyping and interpretation. Sci Transl Med. Apr 24, 2019;11(489):eaat6177. [doi: 10.1126/scitranslmed.aat6177] [Medline: 31019026]
64. Smedley D, Oellrich A, Köhler S, Ruef B, Sanger Mouse Genetics Project, Westerfield M, et al. PhenoDigm: analyzing curated annotations to associate animal models with human diseases. Database (Oxford). 2013;2013:bat025. [FREE Full text] [doi: 10.1093/database/bat025] [Medline: 23660285]
65. Washington NL, Haendel MA, Mungall CJ, Ashburner M, Westerfield M, Lewis SE. Linking human diseases to animal models using ontology-based phenotype annotation. PLoS Biol. Dec 24, 2009;7(11):e1000247. [FREE Full text] [doi: 10.1371/journal.pbio.1000247] [Medline: 19956802]
66. Ayadi A, Birling M, Bottomley J, Bussell J, Fuchs H, Fray M, et al. Mouse large-scale phenotyping initiatives: overview of the European Mouse Disease Clinic (EUMODIC) and of the Wellcome Trust Sanger Institute Mouse Genetics Project. Mamm Genome. Oct 2012;23(9-10):600-610. [FREE Full text] [doi: 10.1007/s00335-012-9418-y] [Medline: 22961258]
67. Bult CJ, Eppig JT, Blake JA, Kadin JA, Richardson JE, Mouse Genome Database Group. The mouse genome database: genotypes, phenotypes, and models of human disease. Nucleic Acids Res. Jan 2013;41(Database issue):D885-D891. [FREE Full text] [doi: 10.1093/nar/gks1115] [Medline: 23175610]
68. Howe DG, Bradford YM, Conlin T, Eagle AE, Fashena D, Frazer K, et al. ZFIN, the Zebrafish Model Organism Database: increased support for mutants and transgenics. Nucleic Acids Res. Jan 2013;41(Database issue):D854-D860. [FREE Full text] [doi: 10.1093/nar/gks938] [Medline: 23074187]

Abbreviations

ACMG: American College of Medical Genetics and Genomics
AI: artificial intelligence

AMP: Association for Molecular Pathology

API: application programming interface

BERT: Bidirectional Encoder Representations from Transformers

CNV: copy number variation

EMR: electronic medical record

GBDT: gradient boosting decision tree

gnomAD: Genome Aggregation Database

GO: Gene Ontology

HGMD: Human Genome Mutation Database

HIPAA: Health Insurance Portability and Accountability Act

HPO: Human Phenotype Ontology

IRB: institutional review board

MViewer: Mutation Viewer

NGS: next-generation sequencing

NLP: natural language processing

NTUH: National Taiwan University Hospital

OMIM: Online Mendelian Inheritance in Man

SNV: single-nucleotide variant

SVM: support vector machine

UMLS: Unified Medical Language System
VCF: variant call format
VEP: Variant Effect Predictor

Edited by A Mavragani; submitted 04.03.22; peer-reviewed by YF Cheng, N Pontikos, C Liuu; comments to author 09.05.22; revised version received 29.07.22; accepted 22.08.22; published 15.09.22

Please cite as:
Huang YS, Hsu C, Chune YC, Liao IC, Wang H, Lin YL, Hwu WL, Lee NC, Lai F
Diagnosis of a Single-Nucleotide Variant in Whole-Exome Sequencing Data for Patients With Inherited Diseases: Machine Learning Study Using Artificial Intelligence Variant Prioritization
JMIR Bioinform Biotech 2022;3(1):e37701
URL: https://bioinform.jmir.org/2022/1/e37701
doi: 10.2196/37701
PMID: