BMC Bioinformatics

SOFTWARE

Open Access

DTreePred: an online viewer based on machine learning for pathogenicity prediction of genomic variants

Daniel Henrique Ferreira Gomes[1,3], Inácio Gomes Medeiros[2], Tirzah Braz Petta[1,4], Beatriz Stransky[1,3] and Jorge Estefano Santana de Souza[1,3*]

*Correspondence:
jorge@imd.ufrn.br

[1] Bioinformatics Postgraduate Program, Metrópole Digital Institute, Federal University of Rio Grande Do Norte, Natal, Rio Grande Do Norte 59078-400, Brazil
[2] Institut Curie, PSL Research University, 26 Rue d'Ulm, 75005 Paris, France
[3] Bioinformatics Multidisciplinary Environment (BioME), Metrópole Digital Institute, Federal University of Rio Grande Do Norte, Natal, Rio Grande Do Norte 59078-400, Brazil
[4] Keck School of Medicine, Department of Translational Genomics, University of Southern California, 1450 Biggy St., Los Angeles, CA 90089, USA

Abstract

Background: A significant challenge in precision medicine is confidently identifying mutations detected in sequencing processes that play roles in disease treatment or diagnosis. Furthermore, the lack of representativeness of single nucleotide variants in public databases and low sequencing rates in underrepresented populations pose defies, with many pathogenic mutations still awaiting discovery. Mutational pathogenicity predictors have gained relevance as supportive tools in medical decision-making. However, significant disagreement among different tools regarding pathogenicity identification is rooted, necessitating manual verification to confirm mutation effects accurately.

Results: This article presents a cross-platform mobile application, DTreePred, an online visualization tool for assessing the pathogenicity of nucleotide variants. DTreePred utilizes a machine learning-based pathogenicity model, including a decision tree algorithm and 15 machine learning classifiers alongside classical predictors. Connecting public databases with diverse prediction algorithms streamlines variant analysis, whereas the decision tree algorithm enhances the accuracy and reliability of variant pathogenicity data. This integration of information from various sources and prediction techniques aims to serve as a functional guide for decision-making in clinical practice. In addition, we tested DTreePred in a case study involving a cohort from Rio Grande do Norte, Brazil. By categorizing nucleotide variants from the list of oncogenes and suppressor genes classified in ClinVar as inexact data, DTreePred successfully revealed the pathogenicity of more than 95% of the nucleotide variants. Furthermore, an integrity test with 200 known mutations yielded an accuracy of 97%, surpassing rates expected from previous models.

Conclusions: DTreePred offers a robust solution for reducing uncertainty in clinical decision-making regarding pathogenic variants. Improving the accuracy of pathogenicity assessments has the potential to significantly increase the precision of medical diagnoses and treatments, particularly for underrepresented populations.

Keywords: Pathogenicity, Mutation, Machine Learning, Viewer, VOUS, Precision medicine

Background

Advancements in cost reduction and efficiency of genetic and genomic tests have led to their widespread adoption in the healthcare sector. These tests are increasingly utilized for diagnosing, monitoring, preventing, and treating various diseases, with a particular emphasis on cancer [1–5]. However, a typical tumor can harbor thousands of mutations, and the genetic heterogeneity can make the unambiguous classification of these alterations even more challenging [6, 7]. Thus, the potential benefit of genetic tests depends on the correct identification and interpretation of pathogenic genetic variants, as they provide a basis for treatment decisions, clinical recommendations, and genetic investigations of the patient's family [8, 9]. A reliable method for identifying the pathogenicity of single nucleotide variants (SNVs) may present significant potential for studying underrepresented cohorts, which can benefit from genetic counseling or precision medicine [10–12]. Families meeting clinical criteria for cancer, in general, could benefit from a more precise assessment of pathogenicity for rare nucleotide variants, usually classified as variants of uncertain significance (VUS) or conflicting data. This identification also provides a basis for effective medicines against diseases through association studies that link genetic variants to drug responses [13, 14]. Accurate pathogenic nucleotide variant identification must be the first step in medicine development, leading to more effective treatments on the basis of a patient's genetic makeup.

Over the last few decades, extensive research has focused on predicting the pathogenicity of genetic nucleotide variants, resulting in the development of numerous tools, databases, and methodologies [15–17]. However, disagreements regarding the pathogenic impact of specific nucleotide variants are more common than expected [18, 19]. For instance, an analysis of nucleotide variant prediction by classical prediction tools revealed discrepancies, such as SIFT classifying the R533H (c.1598G > A) and A4906V (c.14717C > T) variants associated with malignant hyperthermia (MH) as neutral, whereas Polyphen2 and PANTHER classified them as pathogenic [20]. Furthermore, a benchmark study highlighted discrepancies between SIFT and Polyphen2 [21], with an accuracy variation between 70 and 80%.

In this context, machine learning (ML) techniques can help solve such disagreements and predict pathogenicity with greater accuracy on the basis of the volume and variety of available data [20, 22–24]. Recently, our group proposed a decision tree that predicts the pathogenicity of Variants of Uncertain Significance (VUS) [25] and validated it with the ClinVar database [25, 26]. A decision tree is a machine learning algorithm designed to make a sequence of binary decisions on input data, ultimately assigning a classification (or label) to the data. It organizes data in a hierarchical tree structure, making it particularly suitable for problems where input variables are discrete, the target classification is binary, and the model requires clear and interpretable results [27]. The tree presented superior accuracy to several consolidated pathogenicity prediction tools, including SIFT and PolyPhen, by more than ten percentage points and had a competitive advantage over machine learning algorithm models [25]. This algorithm was effectively applied in a case involving a patient with Familial Hemophagocytic Lymphohistiocytosis (FHL) type 5, characterized by a compound heterozygous phenotype in the *STXBP2*

gene, comprising a known pathogenic mutation and a variant of uncertain significance (VUS) not previously linked to FHL [28]. It also enabled the identification of a novel pathogenic *DDB2* variant, NM_000107.3:c.1027G > C, in a case of Brazilian siblings with Xeroderma pigmentosum group E, who developed early-onset melanoma [29]. These applications demonstrate that the decision tree not only accurately classifies VUS but also identifies variants in underrepresented populations.

The use and evaluation of the available nucleotide variant pathogenicity tools can be a challenging task. If the user intends to compare their variant pathogenicity results, they must search for information on different websites, which typically do not provide any guidance for conflicting results. Similarly, building a machine-learning-based pathogenicity prediction tool requires advanced computational skills, a sufficiently large database, and well-curated input data to produce reliable models. Thus, creating a user-friendly application that presents the results of diverse prediction algorithms in one place would be highly convenient for clinicians and researchers, who often use varied online programs and databases for nucleotide variant information to support their clinical decisions [15, 30].

In this article, we present DTreePred, an online viewer for assessing the pathogenicity of nucleotide variants. Users can consult predictions generated by a machine learning-based pathogenicity model, including a decision tree algorithm and 15 machine learning classifiers recently proposed by our group [25], in addition to classical predictors. The target audience comprises researchers and healthcare professionals directly involved in clinical practice, seeking thorough analysis and interpretation of results, especially those applicable to populations underrepresented in public clinical databases. We hope this viewer will help improve knowledge and reduce uncertainties about pathogenic nucleotide variants.

Implementation

ClinVar

ClinVar is a repository created to facilitate the evaluation of nucleotide variants and their associations with phenotypes in a simplified manner [26]. It was established by combining data from various research teams across the globe to examine and validate the feasibility of reaching a consensus on variant analysis outcomes [31]. Its latest version (2024–12-08, available at https://ftp.ncbi.nlm.nih.gov/pub/clinvar/) was acquired and preprocessed as reported in [25], and the application uses it to provide information about the input nucleotide variant passed by the user.

Nucleotide variant annotation and NDamage

Nucleotide variants were annotated for pathogenicity via the dbNSFP database [32] (version 4.8a, available at https://sites.google.com/site/jpopgen/dbNSFP/). Pathogenicity data for nucleotide variants were annotated from twenty-one classical predictor tools (see Supplementary Table 1) and treated as classification variables by the application. This process also obtained allelic frequency information from the Exome Aggregation Consortium (ExAC) database [33] and the 1000Genomes database [34]. The pathogenicity algorithmic predictions were measured as NDamage, a metric

that quantifies the number of classical predictors indicating the nucleotide variants's pathogenicity.

Development of DTreePred

The application implements a decision tree algorithm proposed in [25]. It uses allele frequency data (from ExAC and 1000Genomes) and results from 21 classic predictor tools (see Supplementary Table 1) to determine the final pathogenicity classification. The results are categorized during the annotation phase to generate a meta-prediction, which is subsequently presented as an output by the application, referred to as DTreePred.

Machine learning score

The nucleotide variants of the database ClinVar from the period between 2017 and 2019 were employed to train fifteen pathogenicity classification models, each corresponding to one of the following algorithms: AdaBoost, Bagging, Extra Trees, Random Forest, Logistic Regression, Bernoulli Naive Bayes, Gaussian Naive Bayes, Decision Tree, K-Nearest Neighbors, Multilayer Perceptron, Support Vector Machines (Linear Kernel), Nu-Support Vector Machines, Support Vector Machines, Linear Discriminant Analysis, and Quadratic Discriminant Analysis. These machine learning algorithms' predictions are presented as an output by the application, referred to as Machine Learning Score. The machine learning algorithms implementation used the Scikit-learn framework [35] and the in-house scripts for model training written in Python version 3.10 [36].

In this study, the machine learning algorithms implemented represent five distinct classification strategies. These approaches were designed to capture various data characteristics, with the goal of enhancing the robustness and accuracy of predictions. The strategies include:

- *Ensemble methods*: SKLearn Decision Tree, Extreme Gradient Boosting, Ada Boost, Bagging, Random Forest, and Extra Trees, which combine multiple decision trees to improve precision.
- *Proximity-based models*: K Nearest Neighbors, which classifies data based on Euclidean distances and sample neighborhoods.
- *Hyperplane separation*: Linear Discriminant Analysis, Linear SVMs (with RBF and Linear kernels), Nu-SVC, and Quadratic Discriminant Analysis, which classify samples by dividing the data representation space with hyperplanes.
- *Optimization-based models*: Logistic Regression and Multilayer Perceptron (a neural network with one hidden layer of 100 neurons), which optimize specific mathematical functions to allocate samples to their respective classes.
- *Bayesian probabilistic approaches*: Bernoulli Naive Bayes and Gaussian Naive Bayes, which use conditional probabilities for predictions.

The algorithms were trained and validated using the 10-fold Cross-Validation method. In this approach, the data were divided into 10 subsets; in each iteration, one subset was used for testing, while the remaining nine were used for training. This procedure was repeated for all possible combinations, ensuring a robust evaluation of model performance. Finally, to avoid potential biases, we identified human paralogous genes

using the BioMart [37] tool from *Ensembl* and removed them from the training dataset. A total of 6,363 paralogous genes were selected and used to filter the dataset previously published in [25], resulting in a reduction of approximately 25% in its original size.

The variety of implemented approaches allows for a thorough evaluation of both pathogenic and neutral mutations by leveraging diverse data patterns. Ensemble methods, for example, excel at capturing complex interactions between variables, while probabilistic models provide confidence estimates critical for clinical interpretation. Employing multiple strategies is essential for minimizing biases and enhancing the reliability of predictions in clinical evaluations.

For DTreePred, the integration of different approaches converging on the same pathogenicity classification strengthens the robustness of the results. To this end, we established a meta-prediction threshold, requiring at least six distinct pathogenicity classifications for the Machine Learning Score variable to categorize a mutation as pathogenic. This threshold ensures that the classification reflects input from at least two different classification strategies among the five utilized.

The algorithm implementations, along with the training, validation, and test datasets described in this study, were previously published in [25] and employed in the current implementation of DTreePred.

Application programming interface (back-end)

The Spring Boot extension (https://spring.io/projects/spring-boot) is used to implement the back-end of the system application and is the only component that communicates directly with the application interface. It is responsible for communicating with the dbNSFP database to obtain results from classical predictors [32], which are indexed via the Tabix system [38]. Indexing allows for efficient retrieval of data files related to genomic variants. The API also communicates with a secondary API developed in Flask [39] to obtain prediction results from machine learning algorithms. The

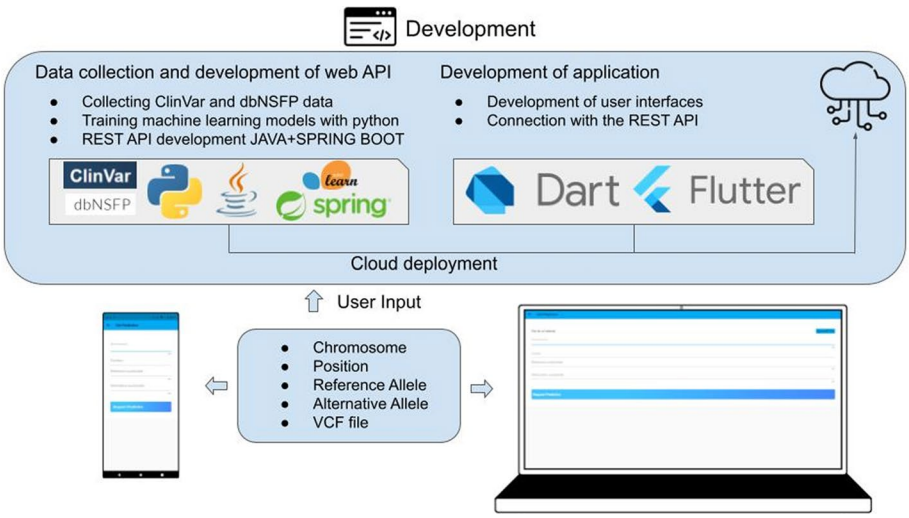

Fig. 1 Development and use of DTreePred. ClinVar and dbNSFP data, as well as clinical and genomic data, are collected, processed, and annotated. Users can text input SNVs or upload SNVs in a tab-separated or Excel file for real-time exploration and interpretation

Table 1 Features Comparison of Variant Viewers

Features	dbNSFP	VEP Web	PROVEAN	CADD	SNPnexus	FAVOR	DTreePred
Presents the results of classic predictors	✓	✓	✓	✓	✓	✓	✓
Stores the inquiries requested by users		✓			*		✓
Human reference genome, version GRCh38—hg38	✓	✓		✓	✓	✓	✓
Human reference genome, version GRCh37—hg19	✓	✓	✓	✓	✓	✓	
Presents predictions based on ML algorithms							✓
Provide decision-making guidance for conflicting data							✓
Presents a native application for mobile devices							✓
Mobile responsive					✓	✓	✓

*Only for 72 h

Java + Spring Boot REST API handles communication with the application and manages user and prediction management tasks. Figure 1 illustrates the API communication architecture.

Interface (front-end)

The mobile application interface was implemented using Flutter version 2.2.1 [40] and Dart version 2.13.1, which are available at https://dart.dev/, to provide an intuitive and user-friendly interface. The design of the interface followed the practices manual provided by Pragmatic Flutter [41], and the architecture followed the business logic components (BLoC) standard [42].

Filters

The DTreePred application incorporates a set of filters to enhance variant analysis and interpretation. These filters include NDmage and Machine Learning Score, which assess the pathogenicity of variants based on multiple predictors and machine learning algorithms. In addition to these scores, DTreePred leverages a curated list of gene groups from various sources, providing context-specific insights. These gene groups enrich variant analysis by offering a context-specific perspective, facilitating a more comprehensive understanding of the potential pathogenicity and relevance of the nucleotide variants analyzed. These include eleven lists related to DNA repair (237 genes), DNA replication (243 genes), oncogenes (71 genes), tumor suppressors (242 genes), KEGG pathways (145 and 36 genes), hallmark cancer genes (1557 genes), hereditary cancer genes (59 genes), and high-functional impact germline polymorphisms (Supplementary Table 2).

General results

DTreePred is a mobile application designed to predict pathogenic nucleotide variants and reconcile predictor discrepancies. We tested DTreePred in an integrity test with 200 variants from ClinVar[26], 172 variants from Leiden Open Variation Database(LOVD,

available at Leiden Open Variation Database https://databases.lovd.nl/shared/variants) [43], whose classification was already known, 557 variants from previously published deep mutational scanning (DMS) experiments [44] and also in a case study composed of a cohort of a population from Rio Grande do Norte, Brazil. Developed to be user friendly, it enables healthcare professionals and researchers to efficiently access a final recommendation on the pathogenicity of a given genetic nucleotide variant. Figure 1 illustrates the development process, features, and workflow. The app is accessible at the following link: https://bioinfo.imd.ufrn.br/dtreepred/.

Upload/Input of genetic loci and data requests for SNVs for clinical interpretation and genetic reports

DTreePred's intuitive interface allows users to request predictions and analyze nucleotide variants. After the user registers or logs in (Fig. 2A), the user will enter the prediction screen. If there are no previous predictions, the user begins by clicking the '+' button (Fig. 2B). To initiate a prediction request, users must enter information such as chromosome number, position, reference allele, and alternate allele of the genetic nucleotide variant (Fig. 2C). By providing these parameters, DTreePred verifies the validity of the chromosome and position. The request is submitted to the back-end server, processed, and redirected to the results page at the front end.

Guide for decision-making

The query results are displayed on a screen with all prediction requests (Fig. 3A). The first highlighted information is the DTreePred result, which indicates whether the nucleotide variant classification is pathogenic or neutral. The screen provides the number of classical predictors (NDamage: out of 21) and machine learning algorithms

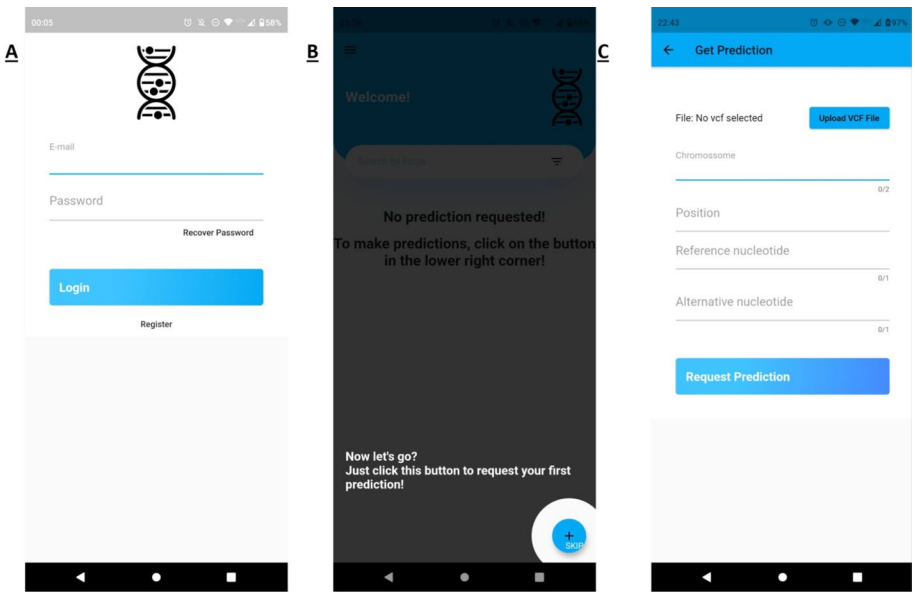

Fig. 2 DTreePred initial screens. **A** App home screen. **B** Prediction display screen without results. If there are no predictions, the user begins by clicking the '+' button. **C** Submission screen. The user must provide either the genomic coordinates (chromosome number, position, reference, and alternate alleles) or a Variant Call Format (VCF file) containing the variants (https://samtools.github.io/hts-specs/VCFv4.5.pdf)

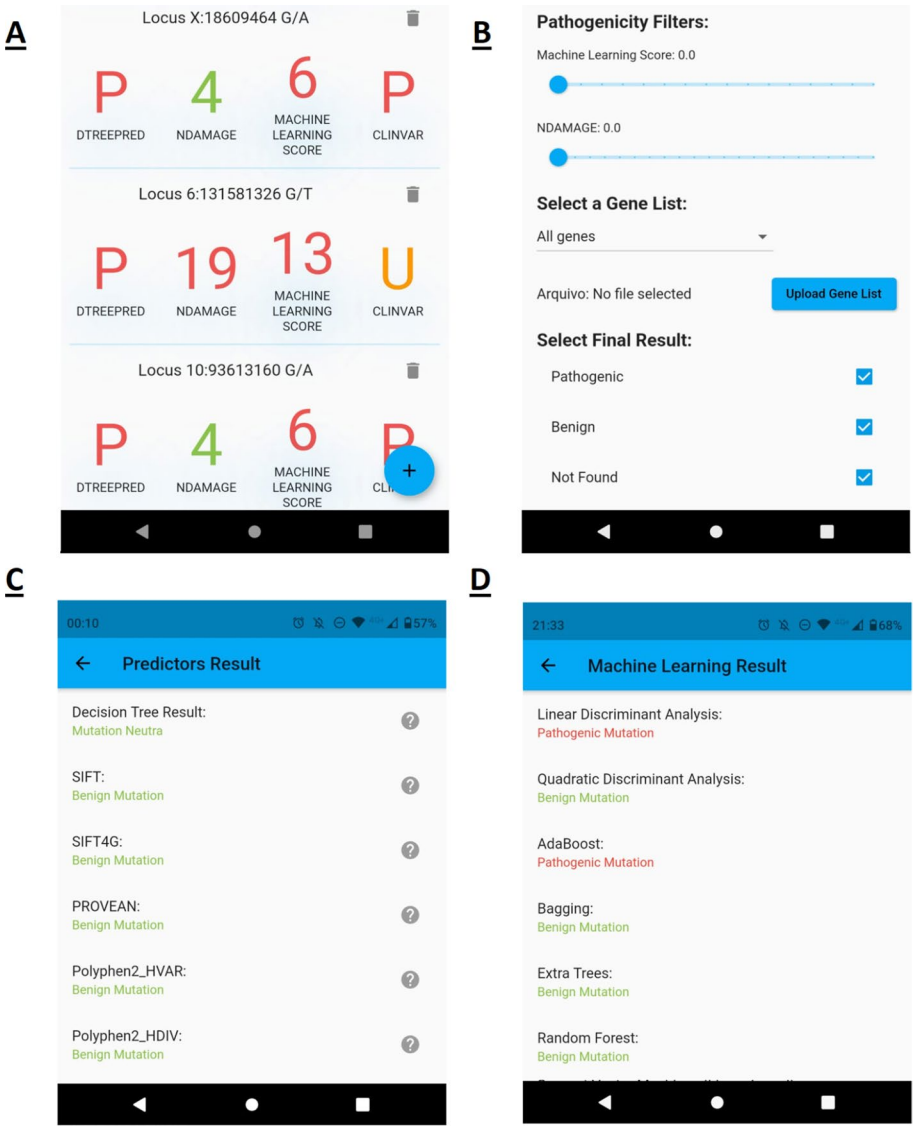

Fig. 3 DTreePred result screens. **A** Overall result, presenting DTreePred displayed as 'N' (neutral) or 'P' (pathogenic), and the number of classical predictors (NDamage) and ML algorithms (Machine Learning Score) that predict a pathogenic variant. If the variant is present in ClinVar, it is also displayed. **B** Advanced search screen to customize the query results. The user can filter the results by prediction scores, gene list, or final prediction. **C** Results from each of the 21 classical predictors. **D** Results from each of the 15 ML algorithms

(Machine Learning Score: out of 15) that reported pathogenic nucleotide variants, emphasizing the consensus among them. Additionally, if the nucleotide variant is present in the ClinVar database, its classification status is also displayed, providing valuable insights from a widely recognized and curated database of genetic variations (Fig. 3A). If it is not practicable to perform the prediction, an error message is exhibited.

These visual elements of DTreePred allow a better assessment of pathogenic nucleotide variants by incorporating multiple methods, predictors, and public data from established databases, generating information of greater accuracy and clinical relevance. It is essential to highlight that the ML algorithms have already demonstrated significant

accuracy compared with classical predictors [25]. Furthermore, these algorithms operate independently and offer a unique and valuable assessment that is not implemented in other tools.

Additional features and analyses provide advanced insights

To facilitate the analysis and improve visualization, we implemented filters to display a specific subset of the requested nucleotide variants (Fig. 3B). In the application, users can select the following options: (i) Machine Learning Score and NDamage: define a minimum score for the respective filters; (ii) List of Genes: chooses a predefined list of genes (Supplementary Table 2) or upload a customized list from a file; (iii) DTreePred: choose the type of decision tree algorithm results displayed on the application (neutral or pathogenic).

The user can check the results of each of the classical predictors on Ndamage (Fig. 3C) or machine learning algorithms on Machine Learning Scores (Fig. 3D) by clicking on their scores. These screens provide a comprehensive view of the assessment of each variant pathogenicity predictor, allowing users to evaluate individual predictions.

Comparison of Variant Viewers

The current pathogenicity predictor tools fulfill the functions of variant prediction and visualization. However, there are various technological limitations, spanning from the absence of adaptability for mobile devices to the lack of incorporation of new prediction techniques. The DTreePred implements and enhances new features compared with various online variant visualization systems, including dbNSFP [32], VEP [45], PROVEAN [46], CADD [47], SNPnexus [48–52] and FAVOR [53]. It offers several functionalities and improvements, such as presenting the results of classic predictors, storing user queries, mobile responsiveness, providing a native mobile application, making predictions based on machine learning algorithms, and offering guidance for decision-making on conflicting data. In Table 1, we present a comparison of the functionalities between DTreePred and online variant viewer systems.

Integrality test

We evaluated the performance of three prediction results from the application (DTreePred, NDamage, and Machine Learning Score) using 200 nucleotide variants with known pathogenicity (100 pathogenic and 100 neutral) sourced from the ClinVar database and not present in the training dataset of the ML algorithms. The confusion matrix presented in Table 2 provides a comprehensive view of each predictor's performance concerning the expected classes (pathogenic and neutral) and the predicted classes. The decision tree output (DTreePred) demonstrated robust capability in identifying variants with potential negative impacts (99% sensitivity) as well as neutral variants (97% sensitivity). The NDamage classifiers exhibited solid proficiency in detecting variants with the possibility of deleterious effects; however, they displayed lower effectiveness for neutral variants (87% success rate for pathogenic variants, 77% for neutral variants). ML algorithms demonstrated balanced performance across both classes (95% success rate for pathogenic variants, 90% for neutral variants).

Table 2 Confusion matrices of the classification performance of the predictors on 200 recent variants (pathogenic and neutral) deposited in ClinVar

		Expected class	
		Pathogenic	***Neutral***
DTreePred (Precision and Recall**: 97.06% and 99.00%)*			
Predicted class	Pathogenic	99	3
	Neutral	1	97
NDamage (Precision and Recall**: 79.09% and 87.00%)*			
Predicted class	Pathogenic	87	23
	Neutral	13	77
ML (Precision and Recall**: 90.48% and 95.00%)*			
Predicted class	Pathogenic	95	10
	Neutral	5	90

*Precision: True positive (pathogenic)/(true positive (pathogenic) + false positive (pathogenic))

**Recall: True positive (pathogenic)/(true positive (pathogenic) + false negative (neutral))

Additionally, to address concerns about potential circular annotation issues within ClinVar, we extended the benchmarking to include an independent dataset derived from the Leiden Open Variation Database (LOVD, version 3). Variants were filtered based on ACMG classification criteria, excluding synonymous variants, variants of uncertain significance (VUS), and those annotated as heterozygous, homozygous, or "different". The results, summarized in Supplementary Table 3, indicate that the DTreePred achieved 94.77% precision (163 correctly classified out of 172 variants available), corroborating its robustness across diverse datasets.

Finally, to address potential annotation issues and biases arising from manual curation, we conducted a benchmarking analysis using a dataset derived from previously published deep mutational scanning (DMS) experiments. As a baseline, we utilized the supplementary data available in the article https://doi.org/10.15252/msb.20199380 [44]. In a second step, to perform double-checking with datasets that were not manually annotated, we also included data from AlphaMissense [54] in our comparison.

We chose to focus on data related to the *BRCA1* gene due to its high relevance in cancer research, in addition to having DMS experiments from two distinct studies [55, 56]. For accuracy calculations, we used the growth rate-based assay [55] as a reference, as it demonstrated superior performance compared to other assays.

Variants with DMS values less than zero were classified as pathogenic, while variants without DMS data in the growth rate-based assay were considered neutral. For these variants without DMS data, we applied an additional filter to exclude those with significant discrepancies in pathogenicity status — specifically, variants labeled as neutral in the original dataset but with more than 10 classical predictors indicating pathogenicity.

The results, summarized in Table 3 and Supplementary Table 4, show that DTreePred achieved the highest accuracy among all evaluated predictors, confirming its robustness across different datasets. Specifically, DTreePred reached an accuracy of 96.95% for the DMS dataset (540 correctly classified out of 557 variants) and 96.72% for the AlphaMissense dataset (530 correctly classified out of 548 variants).

Table 3 Comparison of variant classification accuracy across different predictors in DMS and AlphaMissense datasets

	DMS						AlphaMissense					
Predictors	**Total**	**Hits**	**Errors**	**Accuracy (%)**	**Precision (%)**	**Recal (%)**	**Total**	**Hits**	**Errors**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**
DTreePred	557	540	17	96.95	98.17	87.70	548	530	18	96.72	84.47	97.75
MetaSVM	557	522	35	93.72	82.35	91.80	548	508	40	92.70	93.24	77.53
Envision	557	520	37	93.36	97.83	73.77	548	516	32	94.16	85.39	85.39
LRT	557	518	39	93.00	88.29	80.33	548	518	30	94.53	76.64	92.13
M-CAP	557	517	40	92.82	78.47	92.62	548	496	52	90.51	80.22	82.02
REVEL	557	517	40	92.82	84.17	82.79	548	509	39	92.88	71.55	93.26
CADD	557	514	43	92.28	100.00	64.75	548	523	25	95.44	68.99	100.00
SuSPect	557	492	65	88.33	80.65	61.48	548	514	34	93.80	81.63	88.89
MutationTaster	557	490	67	87.97	65.71	94.26	548	469	79	85.58	68.14	86.52
SNPs&GO	557	482	75	86.54	70.43	66.39	548	500	48	91.24	63.50	97.75
DeepSequence	177	150	27	84.75	94.12	66.67	177	158	19	89.27	52.98	100.00
Ndamage	557	456	101	81.87	54.88	96.72	548	429	119	78.28	42.79	100.00
Polyphen	557	444	113	79.71	52.09	91.80	548	425	123	77.55	41.90	98.88
Machine Learning Score No Homologs	557	408	149	73.25	27.12	13.11	548	423	125	77.19	18.97	12.36
Machine Learning Score	557	403	154	72.35	25.00	13.11	548	418	130	76.28	17.46	12.36
SIFT	557	394	163	70.74	42.49	95.08	548	371	177	67.70	33.46	100.00
PROVEAN	557	368	189	66.07	37.73	84.43	548	362	186	66.06	31.84	95.51
FATHMM	557	361	196	64.81	37.58	91.80	548	332	216	60.58	28.18	92.13

When compared to other meta-predictors, such as MetaSVM, Envision, and REVEL, DTreePred outperformed them in both datasets. For the DMS dataset, MetaSVM achieved an accuracy of 93.72%, Envision 93.36%, and REVEL 92.82%, which are all lower than DTreePred's performance. A similar trend was observed in the AlphaMissense dataset, with MetaSVM and REVEL showing accuracies of 92.70% and 92.88%, respectively, while Envision slightly improved to 94.16%, yet still below DTreePred.

Traditional variant effect predictors, such as CADD, MutationTaster, PolyPhen, and SIFT, exhibited varied performance levels. Among these, CADD demonstrated relatively high accuracy, particularly in the AlphaMissense dataset (95.44%), though still slightly below DTreePred in this specific comparison.

Lower-performing predictors included FATHMM, PROVEAN, and SIFT, with accuracies ranging from 64.81 to 70.74% in the DMS dataset and 60.58 to 67.70% in the AlphaMissense dataset. These results highlight the limitations of older algorithms, especially when handling diverse and complex variant datasets.

Interestingly, the Machine Learning Score without homologs and with homologs models showed moderate accuracy improvements when homologs were excluded, suggesting that paralog filtering may enhance model performance.

Overall, these findings demonstrate that DTreePred not only achieves high accuracy but also exhibits strong generalizability across datasets with different annotation sources, solidifying its potential as a robust tool for variant classification.

A case study

We used our application, DTreePred, to reclassify the variants identified by exome sequencing in a study that included 13 young patients (diagnosed under 40 years old) diagnosed with gastric cancer (GC) from the Northeast region of Brazil [57]. Initially, these patients had negative genetic tests for clinically pathogenic variants (PVs). We categorized the variants considering their impact on specific metabolic pathways associated with DNA repair and DNA replication (RepRep), DNA repair (DnaRep), tumor suppressor genes (Suppressors), protooncogenes (Onco), the CDH1 KEGG pathway (N258_257_61), gastric cancer KEGG (h18 and ko05226), and the list of genes for hereditary cancer from the NCCN (Hereditary_CanGenes). Genetic lists are available in the filter section of DTreePred.

The effectiveness of utilizing pathogenicity predictors in clinical settings hinges on accurately identifying and interpreting potentially pathogenic genetic variants. This comprehension is pivotal for guiding treatment decisions, clinical recommendations, and genetic investigations among other family members of the patient. In the present case study, DTreePred played a crucial role, as over 95% of the variants identified in the oncogenes and tumor suppressor genes lists were classified in the ClinVar database as conflicting data (C), variants of uncertain significance (U), or not found (NF). DTreePred managed to classify all these variants as Neutral or Pathogenic (Fig. 4).

We also analyzed the distribution of genetic effects for variants in oncogenes and tumor suppressor genes. The DTreePred classification proved promising in differentiating patients with long or short overall survival. Patients in stages III or IV presented more pathogenic variants than those in stages II or III (Fig. 5A). The number of pathogenic variants in those genes correlated with advanced tumor stages without

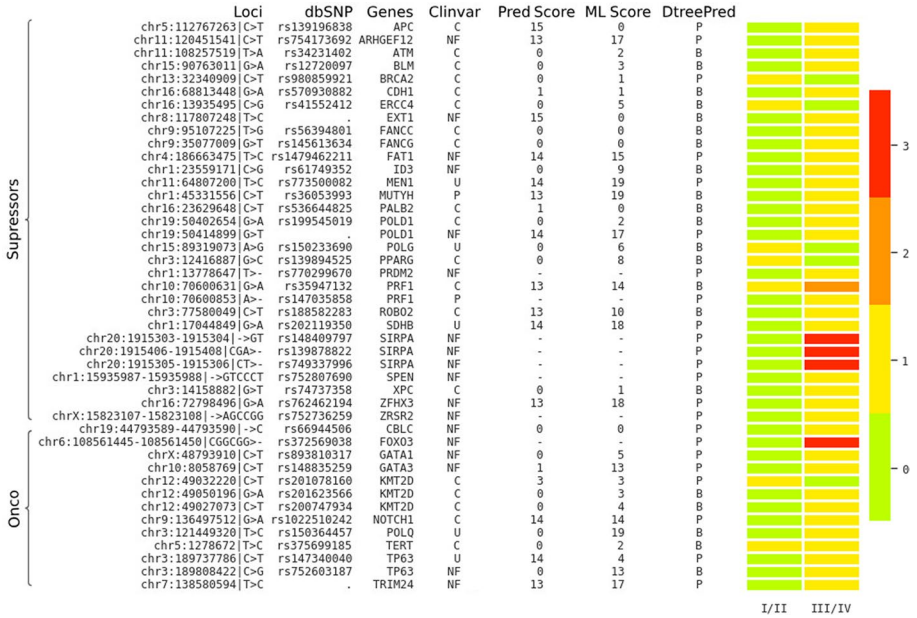

Fig. 4 Pathogenic classification with ClinVar, classical predictions (NDamage), ML algorithms (Machine Learning Score), and DTreePred. In ClinVar, the variant can be classified as neutral (N), pathogenic (P), conflicting data (C), VUS (U), or not found (NF). The color shows the number of patients carrying the variants; I/II or III/IV stages. The predictors and machine learning algorithms predict only nonsynonymous single nucleotide variants (nsSNVs). Out of the 44 variants, 42 had no information in ClinVar

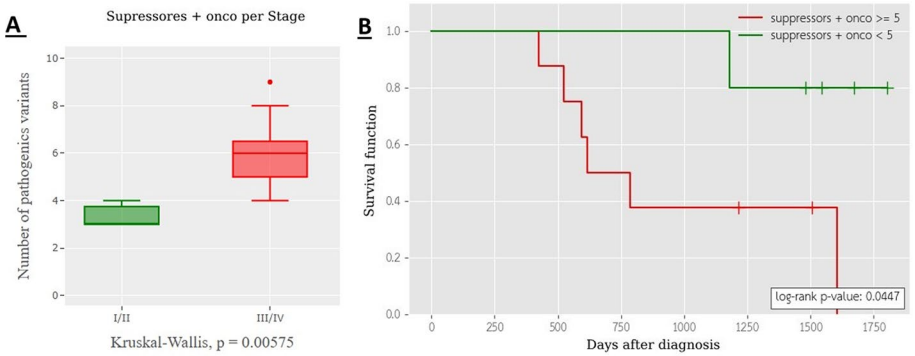

Fig. 5 Clinical impact of > 5 PVs. **A** Distribution of the PVs among GC patients with stage I/II or III/IV (p value = 0.00575). **B** Kaplan–Meier curves showing the associations between overall survival and the number of PVs in protooncogenes and tumor suppressor genes (log-rank test *p* value)

observed associations in other pathways (DTreePred Gene Lists). More importantly, Kaplan–Meier curves revealed a significant association (log-rank test p-value: 0.0447) between overall survival and the number of pathogenic variants in oncogenes and tumor suppressor genes (Fig. 5B). These results highlight the clinical impact of these nucleotide variants on disease progression and patient overall survival.

This example highlights the potential of DTreePred in clinical and scientific routines. It's essential to emphasize that without this tool, this study would not have attained such outcomes; over 95% of the variants were classified as conflicting data, VUS, or not found in the ClinVar database. Additionally, approximately 50% of the variants were

categorized as conflicting or missing data in the tools listed in Table 1, as illustrated in Supplementary Fig. 1. The pathogenicity classification provided by DTreePred, leveraging Decision Tree models and machine learning algorithms, played a pivotal role in conducting a more detailed analysis of the dataset presented in this study.

Discussion

Precision medicine, also known as personalized medicine, is an innovative approach that integrates clinical, genetic, molecular, and environmental information to customize the diagnosis, treatment, and prevention of diseases. A distinguishing characteristic of precision medicine is the ability to identify and interpret genetic variants associated with specific diseases. This represents a significant advancement in clinical practice, enabling the development of highly personalized treatment strategies tailored to individual genetic characteristics, leading to more efficient treatments with fewer side effects. Furthermore, precision medicine advances disease prevention by identifying individuals at greater genetic risk and implementing personalized preventive measures [58].

However, precision medicine faces several significant challenges, and one of the most critical challenges is the accurate interpretation and validation of identified genetic variants. In a thought-provoking comment paper [7], the authors state that "VUSs can cause intense uncertainty for patients and are a vexing challenge for oncologists", emphasizing the clinical consequences of misclassification, such as unnecessary treatments or missed diagnoses. These consequences are related to 3 major circumstances. First, a VUS can be identified in a gene known to confer an inherited predisposition to a cancer, generating implications for family members in determining who should undergo screening. Second, if the VUS is located in a gene associated with a potentially beneficial targeted therapy, it could influence treatment decisions. Lastly, if the VUS turns out to be a driver mutation, there is a risk of overlooking an underlying biological mechanism.

Therefore, the complexity of genetic information and the absence of uniform standards for variant interpretation can interfere with obtaining accurate and clinically meaningful results [18]. Furthermore, the limited availability of high-quality genomic data, especially in underrepresented populations, poses a challenge to the more effective implementation of precision medicine. The lack of genetic diversity in genomic databases can lead to disparities in access to personalized care and the efficacy of treatments across different population groups [59]. We looked at a case study in which 44 different nucleotides were found in oncogenes and suppressor genes. In the ClinVar database, only two of them were clearly classified, leaving the other 42 (approximately 95%) as uncertain.

Despite these difficulties, recent advancements in genomic technology and bioinformatics are propelling ongoing progress in precision medicine. Novel tools and methodologies are emerging to enhance the interpretation of genetic variants, promote diversity within genomic databases, and streamline the integration of genomic data into clinical workflows [60]. The methodologies for variant classification have evolved to encompass a diverse array of data sources, ranging from genomic sequencing data and functional assays to population-level information. By integrating multiple data types, methodologies, and predictors, a more comprehensive assessment of variant

pathogenicity becomes possible, thereby enhancing the accuracy and reliability of predictions.

Both classical predictors and machine learning algorithms play pivotal roles in variant classification. Classical predictors rely on established rules and heuristics to categorize variants on the basis of known patterns [61]; however, with a higher than expected frequency, these predictors yield contradictory information regarding variant pathogenicity. In such instances, meta-predictors, which amalgamate the outcomes of multiple predictors, can help pinpoint the correct classification, especially when the individual predictor results appear discordant [62]. ML algorithms leverage vast genomic datasets to discern intricate patterns and relationships, thereby bolstering prediction accuracy and robustness [63]. Integrated tools that harness these strategies hold tremendous promise in clinical settings.

The effective integration of genomic data into clinical practice is another significant challenge. Healthcare professionals require adequate tools and resources to interpret and apply genetic information in a clinical setting. DTreePred, introduced in this paper, provides comprehensive features tailored to users' needs and consolidates multiple predictions in one place. Notably, the mobile responsiveness of DTreePred further enhances its application in different environments. In practice, a cross-platform application like DTreePred provides significant advantages in clinical settings. For instance, healthcare professionals can seamlessly access the tool on any device—be it a smartphone, tablet, or desktop—ensuring flexibility and convenience. This adaptability enables clinicians to perform variant analysis on the go, whether in a hospital, clinic, or even a remote location, without being constrained by specific operating systems or devices. Furthermore, its compatibility facilitates smooth integration into diverse healthcare workflows, supporting real-time decision-making and advancing personalized medicine.

With a focus on user usability and specific visualization functions, DTreePred implements new methodologies and integrates tools from various sources to construct three scoring systems: Machine Learning Score, NDamage, and DTreePred. It enables users to consult predictions from recently proposed algorithms in [25], in addition to classical predictors, ML algorithms, and the ClinVar database. The tool's role as a guide in decision-making stands out, particularly when the data present contradictions. The Machine Learning Score, incorporates 15 machine learning algorithms; NDamage, 21 classical predictors; and DTreePred, with its decision tree encompassing prediction data and population frequency data, represents an essential tool for precise and comprehensive analysis of variant pathogenicity, offering valuable insights for healthcare professionals and researchers.

We evaluated the performance of the three prediction algorithms using 200 nucleotide variants with known pathogenicity sourced from the ClinVar database. The comprehensive analysis presented in Table 2, through the confusion matrix, indicated that DTreePred outperformed the ML algorithms and NDamage. Notably, the rs78838117 (chr11:2,909,210 G > A) variant, initially classified as neutral by the decision tree but reported as pathogenic in ClinVar (Supplementary Table 5), underscores the complexity of variant interpretation. The high allele frequency in ExAC and 1000 Genomes raises questions about the reliability of the prediction of the ClinVar database,

highlighting the importance of considering factors such as allele frequency, population specificity, and predictive characteristics of machine learning models.

To further assess the performance of DTreePred, we conducted a comparative analysis using variants curated in the LOVD database, applying stringent filters to ensure high-quality benchmarking. The results showed a precision of 94.77%, with 163 out of 172 variants correctly classified. Additionally, to evaluate performance against manually uncurated data, we also conducted a comparison using DMS data. The results showed an accuracy of 96%, with 540 out of 557 variants correctly classified (Supplementary Table 4). This outcome aligns with the performance observed for ClinVar variants and underscores the tool's capacity to generalize effectively across datasets curated under distinct methodologies.

DTreePred also offers a distinct advantage by providing users access to predictions from machine learning algorithms and simultaneous access to multiple predictors in a centralized hub (Fig. 3 and Table 1). Visualizing the construction of each prediction result point-to-point enables users to make different decisions on the basis of their proficiency. Mobile responsiveness also stands out compared to free online viewers, such as the dbNSFP database viewer [32]. While other free viewers have several advantages, such as the Ensembl Variant Effect Predictor [45], which offers interactive annotation, and the PROVEAN web server [46], which provides rapid analysis of protein variants, none of these tools delivers predictions on the basis of models of machine learning such as DTreePred.

Other tools, such as SNPnexus [48] and FAVOR [53], offer comprehensive annotation capabilities and are valuable resources; however, they lack decision-making guidance when predictor results conflict. This limitation can create challenges for users, especially in cases where variant classification is ambiguous. In contrast, DTreePred not only centralizes predictions from multiple models but also provides insights into the decision-making process, equipping users with the ability to resolve conflicts between predictors in a systematic and informed manner.

Understanding the strengths and limitations of these tools allows researchers to select the most appropriate solution for their needs, ensuring reliable and meaningful insights from genomic data.

The simplicity of the application amplifies its usability for various clinical professionals, positioning it as a decision-making guide. This is illustrated by a case study involving 13 patients with gastric cancer in the Northeast region of Brazil, featuring the practical application of DTreePred. By categorizing variants on the basis of specific metabolic pathways and gene lists, the tool assists in genetic investigations, especially when traditional genetic tests are negative. This study showed the effectiveness of DTreePred in classifying variants compared to ClinVar and classical predictors (Fig. 4).

This example also highlights the potential impact of the tool on clinical and scientific routines. The correlation between the number of pathogenic variants and overall patient survival (Fig. 5) points out the clinical relevance of DTreePred's predictions. If DTreePred is used at the time of diagnosis, it would allow the cohort to benefit from genetic counseling or precision medicine. These results underscore the importance of

studying underrepresented populations in which genetic variants are lacking in public databases. Families meeting clinical criteria could benefit from a more precise evaluation of the pathogenicity of rare variants, which are usually classified as VUS or conflicting data. By focusing on these new strategies, we can obtain valuable insights into the genetic landscape of related diseases and develop more comprehensive strategies for diagnosis, prevention, and treatment, showing that the adoption of new tools, such as DTreePred, can change this scenario.

While DTreePred demonstrates robust performance and provides valuable insights into variant classification, there remains significant potential for further enhancement. Key areas for improvement include expanding its scope to analyze variants in regulatory regions, supporting multiple genome versions, and integrating new features and variables into its machine learning models to improve predictive accuracy and adaptability. Also, incorporating feedback mechanisms would allow the application to learn from user inputs and updated datasets, further refining its predictions over time. By addressing these areas, DTreePred has the potential to evolve into a more comprehensive and versatile resource for genomic analysis.

Conclusions

DTreePred represents a significant advancement in the field of precision medicine by addressing one of the core challenges: accurate interpretation of nucleotide variants. Through the integration of machine learning algorithms, classical predictors, and population-level data, this cross-platform mobile application provides healthcare professionals and researchers with a powerful tool for assessing variant pathogenicity. The results of our evaluation demonstrate that DTreePred delivers enhanced accuracy and reliability, outperforming other predictive models in the classification of genetic variants. Its user-friendly interface, mobile responsiveness, and ability to consolidate multiple prediction sources in a centralized hub make it a valuable addition to clinical workflows, especially in environments where quick and reliable decision-making is essential.

The importance of DTreePred extends beyond its technical capabilities. In the context of underrepresented populations, where genetic data is often scarce, DTreePred has the potential to bridge gaps in variant interpretation and enable more equitable access to precision medicine. The practical application of the tool in a case study involving gastric cancer patients from Brazil further underscores its clinical relevance, demonstrating how it can inform genetic counseling and potentially improve patient outcomes. By making advanced bioinformatics accessible, DTreePred not only supports the ongoing evolution of personalized medicine but also highlights the critical need for innovative tools that can adapt to the diverse and complex nature of human genetics.

In conclusion, DTreePred is a pivotal step forward in the integration of genomic data into healthcare, offering a comprehensive and accessible solution for variant classification that holds promise for transforming both research and clinical practice.

Abbreviations

DMS	Deep mutational scanning
ExAC	Exome Aggregation Consortium
GC	Gastric Cancer
ML	Machine Learning
nsSNVs	Nonsynonymous Single Nucleotide Variants
PV	Pathogenic Variants
SNVs	Single Nucleotide Variants
VCF	Variant Call Format
VUS	Variants of Uncertain Significance

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-025-06113-4.

supplementary material 1.

supplementary material 2.

supplementary material 3.

supplementary material 4.

supplementary material 5.

supplementary material 6.

Acknowledgements
We acknowledge the Bioinformatics Multidisciplinary Environment (BioME) at UFRN for the provision of computational resources.

Author contributions
DHFG developed the mobile application and REST API. DHFG and IGM conducted machine learning algorithm training. DHFG and JESS obtained the databases from dbNSFP and Clinvar. IGM, DHFG, BS, TBP, and JESS contributed to data analysis, result interpretation, and manuscript writing. All authors read, commented on, and approved the final version of the manuscript.

Funding
This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Codes: 001 and 88887.687769/2022–00.

Availability of data and materials
Project name: DTreePred. Project home page: https://bioinfo.imd.ufrn.br/dtreepred/. Operating system(s): Platform independent. Programming language: Flutter, Java and Python. Other requirements: Google chrome, Safari, Microsoft Edge or Firefox. License: CC BY-NC 4.0. Any restrictions to use by non-academics: Licence needed. The web application is freely available at https://bioinfo.imd.ufrn.br/dtreepred/ and https://danhfg.github.io/dtreepred/, and the open-source code can be found at https://github.com/Danhfg/Projetos-Bioinfo-front-end/tree/tcc.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 18 October 2024 Accepted: 12 March 2025
Published: 9 April 2025

References

1. Yang T, Li W, Huang T, Zhou J. Genetic testing enhances the precision diagnosis and treatment of breast cancer. Int J Mol Sci. 2023;24:16607.
2. Easwar A, Siddon AJ. Genetic landscape of myeloproliferative neoplasms with an emphasis on molecular diagnostic laboratory testing. Life (Basel). 2021;11:1158.
3. Butz H, Blair J, Patócs A. Molecular genetic testing strategies used in diagnostic flow for hereditary endocrine tumour syndromes. Endocrine. 2021;71:641–52.
4. Grill S, Klein E. Incorporating genomic and genetic testing into the treatment of metastatic luminal breast cancer. Breast Care (Basel). 2021;16:101–7.

5. Premnath N, O'Reilly EM. BReast CAncer (BRCA) gene mutations as an emerging biomarker for the treatment of gastrointestinal malignancies. Chin Clin Oncol. 2020;9:64.
6. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. 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. 2015;17:405–24.
7. Mellgard GS, Atabek Z, LaRose M, Kastrinos F, Bates SE. Variants of uncertain significance in precision oncology: nuance or nuisance? Oncologist. 2024;29:641–4.
8. Nakano Y, Kuiper RP, Nichols KE, Porter CC, Lesmana H, Meade J, et al. Update on recommendations for cancer screening and surveillance in children with genomic instability disorders. Clin Cancer Res. 2024;30:5009–20.
9. Kwong A, Ho CYS, Shin VY, Au CH, Chan T-L, Ma ESK. How does re-classification of variants of unknown significance (VUS) impact the management of patients at risk for hereditary breast cancer? BMC Med Genomics. 2022;15:122.
10. Kobayashi Y, Chen E, Facio FM, Metz H, Poll SR, Swartzlander D, et al. Clinical variant reclassification in hereditary disease genetic testing. JAMA Netw Open. 2024;7:e2444526.
11. Manotas MC, Rivera AL, Sanabria-Salas MC. Variant curation and interpretation in hereditary cancer genes: An institutional experience in Latin America. Mol Genet Genomic Med. 2023;11:e2141.
12. Appelbaum PS, Burke W, Parens E, Zeevi DA, Arbour L, Garrison NA, et al. Is there a way to reduce the inequity in variant interpretation on the basis of ancestry? Am J Hum Genet. 2022;109:981–8.
13. Singh DB. The impact of pharmacogenomics in personalized medicine. Adv Biochem Eng Biotechnol. 2020;171:369–94.
14. Hockings JK, Pasternak AL, Erwin AL, Mason NT, Eng C, Hicks JK. Pharmacogenomics: an evolving clinical tool for precision medicine. Cleve Clin J Med. 2020;87:91–9.
15. Walters-Sen LC, Hashimoto S, Thrush DL, Reshmi S, Gastier-Foster JM, Astbury C, et al. Variability in pathogenicity prediction programs: impact on clinical diagnostics. Mol Genet Genomic Med. 2015;3:99–110.
16. Garcia FA, Andrade ES, Palmero EI. Insights on variant analysis in silico tools for pathogenicity prediction. Front Genet. 2022;13:1010327.
17. Lin Y-J, Menon AS, Hu Z, Brenner SE. Variant Impact Predictor database (VIPdb), version 2: trends from three decades of genetic variant impact predictors. Hum Genomics. 2024;18:90.
18. Amendola LM, Jarvik GP, Leo MC, McLaughlin HM, Akkari Y, Amaral MD, et al. Performance of ACMG-AMP variant-interpretation guidelines among nine laboratories in the clinical sequencing exploratory research consortium. Am J Hum Genet. 2016;99:247.
19. Stella S, Vitale SR, Massimino M, Martorana F, Tornabene I, Tomarchio C, et al. In silico prediction of BRCA1 and BRCA2 variants with conflicting clinical interpretation in a cohort of Breast Cancer patients. Genes (Basel). 2024;15:943.
20. Schiemann AH, Stowell KM. Comparison of pathogenicity prediction tools on missense variants in RYR1 and CACNA1S associated with malignant hyperthermia. Br J Anaesth. 2016;117:124–8.
21. Suybeng V, Koeppel F, Harlé A, Rouleau E. Comparison of pathogenicity prediction tools on somatic variants. J Mol Diagn. 2020;22:1383–92.
22. Lai C, Zimmer AD, O'Connor R, Kim S, Chan R, van den Akker J, et al. LEAP: using machine learning to support variant classification in a clinical setting. Hum Mutat. 2020;41:1079–90.
23. Nicora G, Zucca S, Limongelli I, Bellazzi R, Magni P. A machine learning approach based on ACMG/AMP guidelines for genomic variant classification and prioritization. Sci Rep. 2022;12:2517.
24. Katsonis P, Wilhelm K, Williams A, Lichtarge O. Genome interpretation using in silico predictors of variant impact. Hum Genet. 2022;141:1549–77.
25. do Nascimento PM, Medeiros IG, Falcão RM, Stransky B, de Souza JES. A decision tree to improve identification of pathogenic mutations in clinical practice. BMC Med Inform Decis Mak. 2020;20:52.
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. 2014;42:D980-5.
27. Mitchell T. Machine Learning. New York, NY: McGraw-Hill Professional; 1997.
28. de Souza Timoteo AR, Pinheiro de Almeida IC, Yurchenko AA, de Miranda Henriques SR, de Souza SP, Rajabi F, et al. Brazilian XP-E siblings carrying a novel DDB2 variant developed early-onset melanoma: a case report. BMC Med Genomics. 2023. https://doi.org/10.1186/s12920-023-01622-8.
29. Dantas VM, Valle CT, de Oliveira RP, Bezerra MTAL, do Amaral CT, Brandão RAS, et al. Germline compound heterozygous variants identified in the STXBP2 gene leading to a familial hemophagocytic lymphohistiocytosis type 5: A case report. Front Pediatr. 2021. https://doi.org/10.3389/fped.2021.633996.
30. Iversen J, Eierman M. Learning Mobile App Development: A Hands-on Guide to Building Apps with IOS and Android. Addison-Wesley; 2014.
31. Harrison SM, Riggs ER, Maglott DR, Lee JM, Azzariti DR, Niehaus A, et al. Using ClinVar as a resource to support variant interpretation. Curr Protoc Hum Genet. 2016. https://doi.org/10.1002/0471142905.hg0816s89.
32. Liu X, Li C, Mou C, Dong Y, Tu Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. Genome Med. 2020;12:103.
33. Karczewski KJ, Weisburd B, Thomas B, Solomonson M, Ruderfer DM, Kavanagh D, et al. The ExAC browser: displaying reference data information from over 60 000 exomes. Nucleic Acids Res. 2017;45:D840–5.
34. Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526:68–74.
35. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. J Mach Learn Res. 2011;12:2825–30.
36. Van Rossum G, Drake FL. Python 3 Reference Manual: (Python Documentation Manual Part 2). CreateSpace Independent Publishing Platform; 2009.
37. Steffen D, Wolfgang H (2017) biomaRt. Bioconductor
38. Li H. Tabix: fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics. 2011;27:718–9.
39. Grinberg M. Flask Web Development. "O'Reilly Media, Inc."; 2018.

40. Napoli ML. Introducing flutter and getting started. Beginning Flutter: A Hands-on Guide to App Development. Wiley, Indianapolis, USA; 2019.
41. Tyagi P. Pragmatic Flutter: Building Cross-Platform Mobile Apps for Android, IOS. Web and Desktop: CRC Press; 2021.
42. GitHub - jacobaraujo7/bloc-pattern: Apenas um package com bases para implantar o Bloc no seu Código. GitHub. https://github.com/jacobaraujo7/bloc-pattern. Accessed 18 Jan 2024.
43. Fokkema IFAC, Kroon M, López Hernández JA, Asscheman D, Lugtenburg I, Hoogenboom J, et al. The LOVD3 platform: efficient genome-wide sharing of genetic variants. Eur J Hum Genet. 2021;29:1796–803.
44. Livesey BJ, Marsh JA. Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations. Mol Syst Biol. 2020;16:e9380.
45. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17:122.
46. Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics. 2015;31:2745–7.
47. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019;47:D886–94.
48. Oscanoa J, Sivapalan L, Gadaleta E, Dayem Ullah AZ, Lemoine NR, Chelala C. SNPnexus: a web server for functional annotation of human genome sequence variation (2020 update). Nucleic Acids Res. 2020;48:W185–92.
49. Dayem Ullah AZ, Oscanoa J, Wang J, Nagano A, Lemoine NR, Chelala C. SNPnexus: assessing the functional relevance of genetic variation to facilitate the promise of precision medicine. Nucleic Acids Res. 2018;46:W109–13.
50. Dayem Ullah AZ, Lemoine NR, Chelala C. A practical guide for the functional annotation of genetic variations using SNPnexus. Brief Bioinform. 2013;14:437–47.
51. Dayem Ullah AZ, Lemoine NR, Chelala C. SNPnexus: a web server for functional annotation of novel and publicly known genetic variants (2012 update). Nucleic Acids Res. 2012;40:W65-70.
52. Chelala C, Khan A, Lemoine NR. SNPnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms. Bioinformatics. 2009;25:655–61.
53. Zhou H, Arapoglou T, Li X, Li Z, Zheng X, Moore J, et al. FAVOR: functional annotation of variants online resource and annotator for variation across the human genome. Nucleic Acids Res. 2023;51:D1300–11.
54. Cheng J, Novati G, Pan J, Bycroft C, Žemgulytė A, Applebaum T, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science. 2023;381:eadg7492.
55. Findlay GM, Daza RM, Martin B, Zhang MD, Leith AP, Gasperini M, et al. Accurate classification of BRCA1 variants with saturation genome editing. Nature. 2018;562:217–22.
56. Starita LM, Young DL, Islam M, Kitzman JO, Gullingsrud J, Hause RJ, et al. Massively parallel functional analysis of BRCA1 RING domain variants. Genetics. 2015;200:413–22.
57. CGA-IGC (2023) Abstracts Fam. Cancer. 2024;23:41–107
58. Pastorino R, Loreti C, Giovannini S, Ricciardi W, Padua L, Boccia S. Challenges of prevention for a sustainable personalized medicine. J Pers Med. 2021;11:311.
59. Landry LG, Ali N, Williams DR, Rehm HL, Bonham VL. Lack of diversity in genomic databases is a barrier to translating precision medicine research into practice. Health Aff. 2018;37:780–5.
60. Manolio TA, Rowley R, Williams MS, Roden D, Ginsburg GS, Bult C, et al. Opportunities, resources, and techniques for implementing genomics in clinical care. Lancet. 2019;394:511–20.
61. Cristianini N, Scholkopf B. Support vector machines and kernel methods: the new generation of learning machines. AIMag. 2002;23:31–31.
62. Forero DA. Bioinformatics and Human Genomics Research. Taylor & Francis Group; 2021.
63. Sarker M. Revolutionizing healthcare: the role of machine learning in the health sector. JAIGS. 2024;2:36–61.

Publisher's Note