PLOS ONE

RESEARCH ARTICLE

WilsonGenAI a deep learning approach to classify pathogenic variants in Wilson Disease

Aastha Vatsyayan[1,2‡], **Mukesh Kumar**[1,2‡], **Bhaskar Jyoti Saikia**[1,2], **Vinod Scaria**[1,2¤*], **Binukumar B. K.**[1,2*]

1 CSIR Institute of Genomics and Integrative Biology (CSIR-IGIB), Delhi, India, **2** Academy of Scientific and Innovative Research (AcSIR), Ghaziabad, India

¤ Current address: Vishwanath Cancer Care Foundation, Mumbai, India
‡ AV and MK contributed equally and would like to be known as joint first authors.
* vinods@igib.in (VS); binukumar@igib.in (BBK)

OPEN ACCESS

Citation: Vatsyayan A, Kumar M, Saikia BJ, Scaria V, B. K. B (2024) WilsonGenAI a deep learning approach to classify pathogenic variants in Wilson Disease. PLoS ONE 19(5): e0303787. https://doi.org/10.1371/journal.pone.0303787

Editor: Muhammad Salman Bashir, King Fahad Medical City, SAUDI ARABIA

Received: November 14, 2023

Accepted: May 1, 2024

Published: May 17, 2024

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0303787

Data Availability Statement: All relevant data are within the manuscript and its Supporting information files.

Funding: This work was supported by the Council of Scientific and Industrial Research (CSIR)

Abstract

Background

Advances in Next Generation Sequencing have made rapid variant discovery and detection widely accessible. To facilitate a better understanding of the nature of these variants, American College of Medical Genetics and Genomics and the Association of Molecular Pathologists (ACMG-AMP) have issued a set of guidelines for variant classification. However, given the vast number of variants associated with any disorder, it is impossible to manually apply these guidelines to all known variants. Machine learning methodologies offer a rapid way to classify large numbers of variants, as well as variants of uncertain significance as either pathogenic or benign. Here we classify *ATP7B* genetic variants by employing ML and AI algorithms trained on our well-annotated WilsonGen dataset.

Methods

We have trained and validated two algorithms: TabNet and XGBoost on a high-confidence dataset of manually annotated, ACMG & AMP classified variants of the *ATP7B* gene associated with Wilson's Disease.

Results

Using an independent validation dataset of ACMG & AMP classified variants, as well as a patient set of functionally validated variants, we showed how both algorithms perform and can be used to classify large numbers of variants in clinical as well as research settings.

Conclusion

We have created a ready to deploy tool, that can classify variants linked with Wilson's disease as pathogenic or benign, which can be utilized by both clinicians and researchers to better understand the disease through the nature of genetic variants associated with it.

[IndiGenApp Grant and OLP2301]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abbreviations: ACMG-AMP, American College of Medical Genetics and Genomics and the Association of Molecular Pathologists; MCC, Matthews Correlation Coefficient; NPV, Negative Predictive Value; PPV, Positive Predictive Value; SDM, Site-Directed Mutagenesis.

Introduction

Wilson's Disease (WD) is a rare autosomal recessive disorder characterized by the presence of pathogenic mutations in the copper-transporting *ATP7B* gene. Located on chromosome 13q14.2, *ATP7B* spans 21 exons, encoding a 1465-amino-acid copper-transporting ATPase [1]. Altered gene function in WD results in copper accumulation in the liver and brain, leading to impaired functions and movement disorders. WD patients exhibit pathogenic mutations causing reduced serum holo-ceruloplasmin production. Excessive copper deposition induces oxidative stress, contributing to clinical problems like cirrhosis and fulminant hepatitis. Neurological complications arise from copper deposits in specific brain regions, leading to movement disorders and associated symptoms [2]. This complex interplay of genetic factors and copper metabolism underscores the multisystemic impact of WD.

WD is an underdiagnosed and treatable genetic condition with an estimated worldwide prevalence of around 13.9 per 100,000, derived from known pathogenic variants [3]. Several recent publications have highlighted an estimated carrier frequency of 1 in 90 individuals [4–6]. The known prevalence and carrier frequency of WD however, are confined to a few specific populations [7, 8] while in large populations like India, they remain unexplored. This opens up a unique opportunity to understand the genetic architecture of the disease in populations rich in genetic diversity such as India.

The recent availability of a framework for the interpretation of pathogenicity of genetic variants put forward by the American College of Medical Genetics and Genomics and the Association of Molecular Pathologists (ACMG & AMP) has opened up a unique opportunity to create a standardised system for interpretation of genetic variants for clinical diagnosis and genetic counselling. To assist in a better understanding of variant pathogenicity, our group has recently put together one of the most comprehensive collections of genetic variants classified according to the ACMG & AMP Guidelines [9], in form of the WilsonGen database, a robust compilation of all publicly reported *ATP7B* variants exhaustively collected from literature and across 9 large databases [10], making it the largest, most comprehensive database of it's kind, to the best of our knowledge.

While classification according to the ACMG & AMP guidelines is time-consuming and at times limited by literature and experimental evidence to confirm the pathogenicity, a number of variants remain unclassified as variants of uncertain significance (VUS). This significantly impacts the ability to classify variants, especially from unique population groups and rare variants identified from patient cohorts.

The advent of machine learning approaches in clinical medicine have accelerated the ability to analyse and interpret medical data and have been extensively used in a number of scenarios, including the rapid classification of large numbers of variants. The widespread application of such approaches in genomics however, has been limited by the lack of gold-standard datasets for training. The availability of WilsonGen database thus provides a unique opportunity in this aspect.

Here, we describe a machine learning approach trained on a gold-standard ACMG-classified variant dataset for pathogenicity in the *ATP7B* gene for accurate classification of variants. We also use the approach for reclassification of VUS variants in public datasets so as to enable quick variant interpretation in clinical and research settings. To the best of our knowledge, ours is the only approach based on a manually ACMG-classified dataset, dedicated specifically to WD variants. A public implementation of the algorithm is available at: https://github.com/aastha-v/WilsonGenAI.

Materials and methods

Datasets

The variants and their pathogenicity ascertained according to the ACMG and AMP guidelines and available in the WilsonGen database were taken up for analysis. This dataset contained a total of 1458 genetic variants manually classified according to the ACMG & AMP guidelines. Non-exonic variants were removed due to lack of sufficient training data, as were VUS variants. This resulted in a variant dataset of 723 unique variants, out of which 410 were annotated as pathogenic, 167 as likely pathogenic, 9 as benign and 137 as likely benign. Fig 1 offers an overview of our entire workflow.

Variant parameters

The variant VCF was run through the ANNOVAR [11] tool, which annotated the variants with allele frequencies (AF) from three global population and subpopulation datasets: GnomAD [12], 1000Genomes [13] and GME [14]. We further added the position of the first

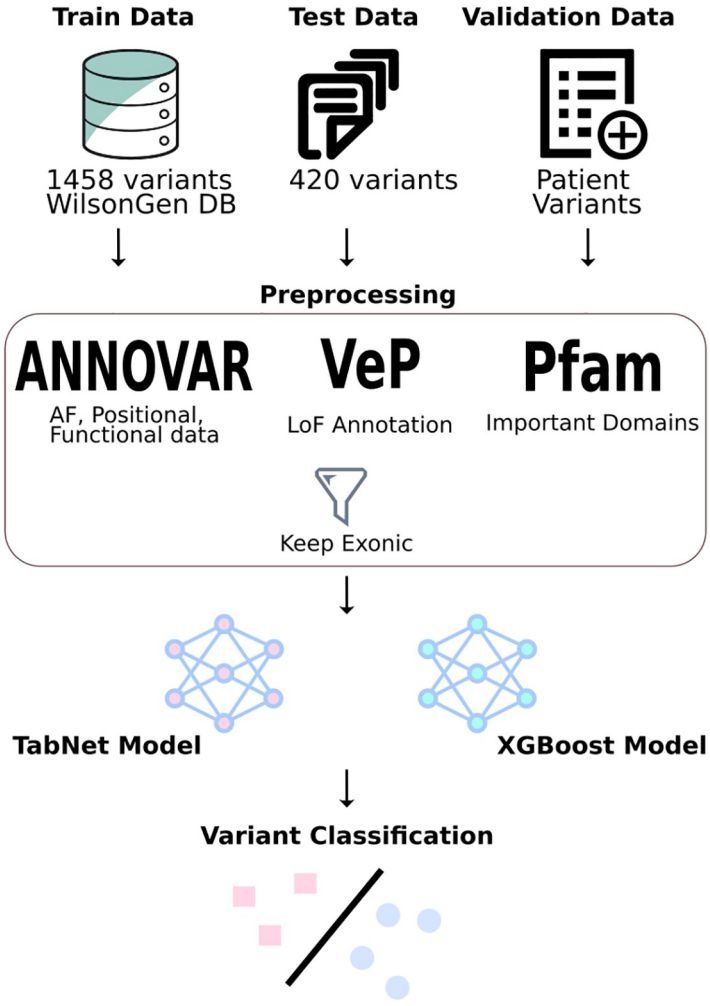

Fig 1. The overview of the workflow followed for model development and variant classification with TabNet and XGBoost models.

https://doi.org/10.1371/journal.pone.0303787.g001

amino acid change for each variant as "Start_pro", which would offer positional data along with "Start" which marked the position of nucleotide change. Further, we added the categorical attributes "Pfam_imp_domain" that would indicate whether the variant overlapped with an important protein domain based on the Pfam database, as well as "LoF_HC_Canonical" categorical attribute which would indicate if the variant was a high-confidence loss-of-function variant present in a canonical transcript. The exonic function (e.g. frameshift insertion/deletion. Stopgain/Startloss etc.) was further encoded into numbers. Feature selection was performed manually with all features with missing values that exceeded 80% for each class being removed. Finally, all /likely pathogenic variants were encoded as "1" and all benign/likely benign variants as "0". A total of 73 attributes were thus obtained and are detailed in S1 Table.

AI models

For our analysis, we considered two state-of-the-art deep and machine learning models, namely TabNet and XGBoost, to train on the ACMG-classified gold standard dataset.

We had previously utilized the Weka suite [15] to test the performance of several algorithms including NaiveBayes, SMO, J48, and RandomForest on the dataset available to us then, comprising of 725 variants split into 70% train—30% test datasets. Traditionally, tree ensemble models are recommended for classification and regression problems with tabular data [16]. Our results proved to be in concordance, since RandomForest and J48 outperformed others. We thus chose to work with the XGBoost classifier, which is one of the most widely used gradient-boosted decision trees, especially for tabular datasets. XGBoost is reported to perform faster and better than other models such as RandomForest for missing data, and with class imbalanced datasets. It also has in-built regularization which prevents overfitting, which models like RandomForest can be prone to. The XGBoost algorithm creates decision trees in sequential form, wherein increased weights of incorrectly predicted variables are fed into the next tree. The algorithm has been created to handle sparse data effectively, which mirrors real-world situations where data is often found to be missing or containing frequently repeating values.

Additionally, we chose to also utilize the novel deep learning neural network TabNet [17], which was specially created for tabulated datasets. TabNet has been reported to outperform tree methods including XGBoost for certain tabular datasets [18]. Unlike other deep learning models, TabNet mimics the learning of decision trees, through the use of its transformer architecture, enabling the model to quickly decipher complex data patterns. TabNet uses sequential attention to choose features at each decision step. Feature selection is done instance-wise, i.e. it could be different for each variant in the training dataset. To the best of our knowledge, this is the first implementation of TabNet for the classification of variants based on their pathogenicity.

Since the performance of the two models with respect to each other seems to vary based on datasets used [16], we decided to include results from both models for assessment.

Hyperparameter selection and cross validation

Our models were run with different input parameters until convergence. The best performing model by accuracy was taken up. The PyTorch [19] implementation of Google's TabNet was used for model creation, while Anaconda [20] was used to enable the use of Scikit-learn, Pandas, Matplotlib and Seaborn to enable analysis and visualisation for both models.

For TabNet, SimpleImputer was used to replace missing data with a constant value. Further, the model's mask_type parameter was set to 'entmax', which showed a better overall performance than the default 'sparsemax'. The 'weights' parameter was set to '1' to address class imbalance, while the batch size was set at the maximum recommended 10% of the total data

size at 72. A maximum of 1000 epochs were allowed with a patience (i.e. the number of epochs to wait for improvement before terminating the training run) of 100.

For XGBoost, the hyper-parameters were selected and evaluated using a 5-fold cross validation (CV) approach. A randomized search on the hyperparameters was performed using RandomizedSearchCV (CV = 5). Class imbalance was corrected using the scale_pos_weight parameter set at 3.95. The following hyper-parameters were finally used (Table 1):

The mean cross_val_score function (CV = 10) was used to test model performance for both models across multiple test/train splits. Several models with and without the hyperparameters determined during tuning, were tested for performance using accuracy metrics described below. The best performing model was then selected.

Independent validation dataset

An additional number of 420 variants were curated from published literature not included in the WilsonGen database till 2022. The variants were classified according to the ACMG & AMP guidelines as described previously. The dataset comprised of 31 variants which were annotated as pathogenic, 29 which were likely pathogenic, 96 were likely benign, and the remaining variants were classified as VUS. Thus we had a total of 156 variants in our independent test dataset.

Accuracy estimates

The following accuracy estimates were used to evaluate the performance of the models: a) Sensitivity b) Specificity c) Accuracy d) Positive Predictive Value (PPV) e) Negative Predictive Value (NPV), and f) Matthews Correlation Coefficient (MCC). All data used in this study is freely accessible at: https://clingen.igib.res.in/WilsonGen/ The source code of both our models is available at https://github.com/aastha-v/WilsonGenAI. The models have been standardized on Ubuntu 18 LTS. The instructions and code for the preprocessing pipeline, variant classification through our models, as well as for generating one's own models are also freely included.

Patient data validation

Generating variants and functional validation. *ATP7B* plasmid (pLB1080; Addgene) was subjected to site-directed mutagenesis (SDM) according to the manufacturer's instruction (Agilent, 200522) using the set of primers shown in Table 2.

To understand the impact of WT (wild type) *ATP7B* and its protein mutants, knock-out HepG2 cells were cultured under the standard conditions. Different plasmids were transfected using lipofectamine-3000 (Thermo Scientific, L3000008). Post 24 hours, cells were treated

Table 1. Model hyperparameters used for the XGBoost model.

Hyperparameter	Value	Hyperparameter	Value	Hyperparameter	Value
base_score	0.5	gpu_id	-1	min_child_weight	1
booster	gbtree	grow_policy	depthwise'	missing	nan
callbacks	None	importance_type	None	monotone_constraints	()
colsample_bylevel	1	interaction_constraints	'	n_estimators	50
colsample_bynode	1	learning_rate	0.25	n_jobs	0
colsample_bytree	0.9	max_bin	256	num_parallel_tree	1
early_stopping_rounds	None	max_cat_to_onehot	4	predictor	auto
enable_categorical	FALSE	max_delta_step	0	random_state	0
eval_metric	None	max_depth	6	reg_alpha	0
gamma	0.2	max_leaves	0	reg_lambda	1

https://doi.org/10.1371/journal.pone.0303787.t001

Table 2. Primers used in site-directed mutagenesis.

Variants	Forward Primer (5'—3')	Reverse Primer (5'—3')
c.2564C>T (S855F)	CTCCTGTGATGAGGAACTCATCAGCCATGGTATT	AATACCATGGCTGATGAGTTCCTCATCACAGGAG
c.813C>A (C271X)	GCCTCCGCAGTCTCCACCACAGCCA	TGGCTGTGGTGGAGACTGCGGAGGC

https://doi.org/10.1371/journal.pone.0303787.t002

with 500 μM CuCl2 for 6 hours and replaced with fresh media. After 18 hours, spent media was collected to estimate the exported copper using the manufacturer's protocol (Sigma, MAK127). The colorimetric data of the assay was analyzed using an unpaired-t-test, with a p-value$<$0.05 considered statistically significant for all three sets of experiments.

Results

Both models performed best with a 70–30% train-test split. The TabNet model additionally split the 30% test set into 50% train and 50% validation subsets during training.

Accuracy estimates

TabNet. Although the model was set to run at a maximum of 1000 epochs, it stopped the training at 187th epoch with the best accuracy of 99% on the validation and 97.24% on the test sets respectively. The overall MCC was 0.92. The Precision, Recall and F1 scores are shown in S2 Table. S1 Fig shows the accuracy and Fig 2 the Area Under the Curve (AUC) plot; the receiver operating characteristic curve (ROC) was 0.996. Further, S2 Fig shows the confusion matrix for our test data; out of 109 variants taken as part of the 50% test subset data, the model accurately predicted 84 as pathogenic and 22 as benign. The Precision-Recall curve is shown in S3 Fig, with the overall area under the precision-recall curve (AUPRC) determined to be 1. The model learning rate and loss are plotted in S4 Fig. Additionally, the model Specificity, and its Negative Predictive Value (NPV) were both 1.

XGBoost. The XGBoost model had an overall accuracy on the test set of 0.986175, AUC 0.9926 and MCC of 0.952773. Fig 2 shows the AUC plot, while S2 Fig depicts the confusion matrix. The Precision, Recall and F1 scores are shown in S2 Table. The Precision-Recall curve is shown in S3 Fig, with the overall AUPRC determined to be 1. Additionally, the model Specificity was 0.989, and its NPV was 1.

Validation in an independent set of variants classified according to the ACMG & AMP guidelines

After removing all non-exonic variants, we had a total of 96 benign/likely benign variants clubbed together as benign, and 60 pathogenic/likely pathogenic variants clubbed together as "pathogenic". Upon running our models on the data, the TabNet model accurately classified all correctly, while XGBoost correctly classified 60 variants as pathogenic and 95 as benign, as shown in the confusion matrices in S5 Fig. Scatterplots of class probability vs the actual ACMG class for each model across all 156 variants are shown in S6 and S7 Figs for TabNet and XGBoost respectively.

Comparison with CADD

Both our models performed better than CADD, which only had scores for 53 out of the 156 variants included in the independent ACMG-qualified test set. TabNet had an overall accuracy on the test set of 1, and XGBoost of 0.9935, while CADD only had an overall accuracy of

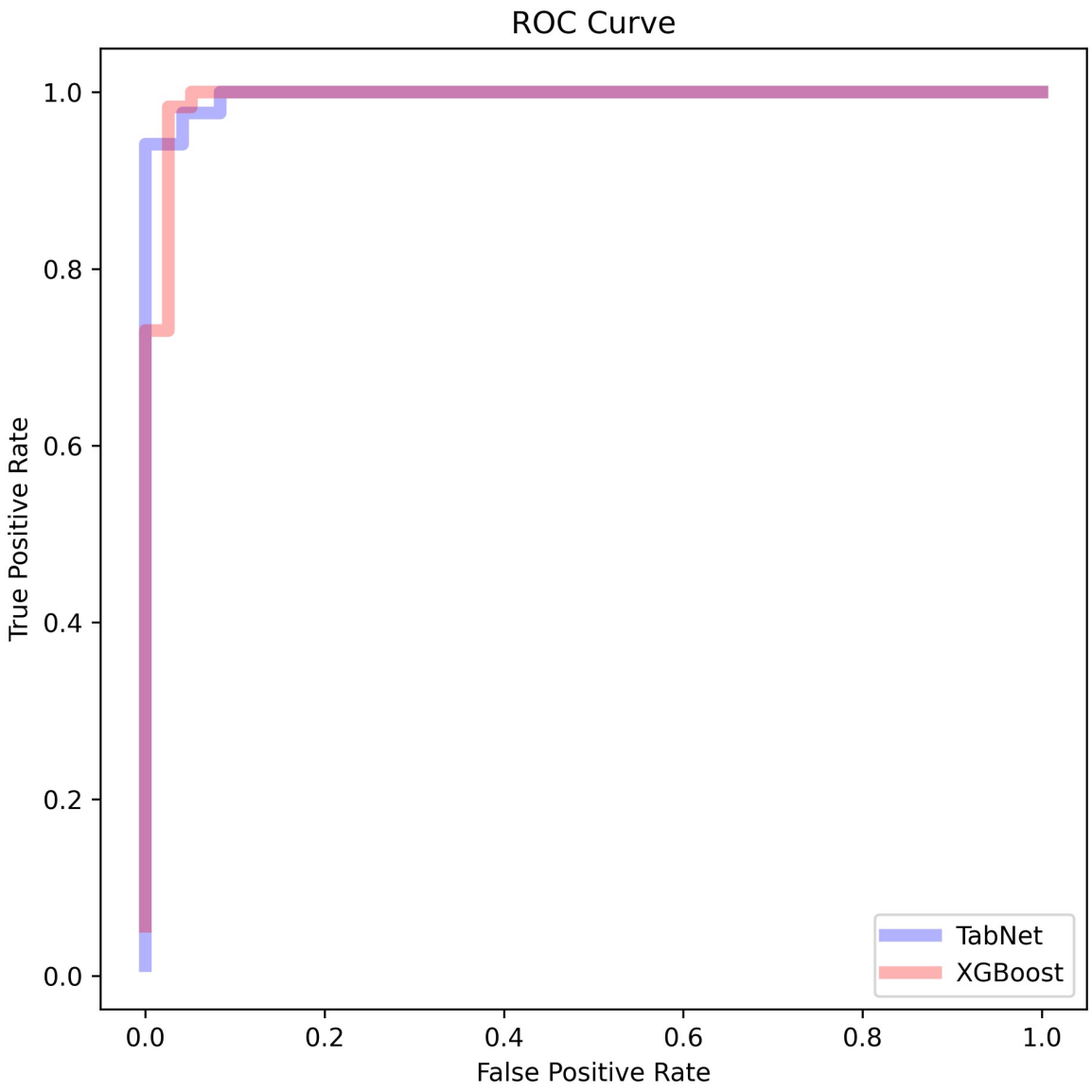

Fig 2. The receiver operating characteristic curve for (A) the TabNet and (B) the XGBoost model.

https://doi.org/10.1371/journal.pone.0303787.g002

0.9811 on the limited number of variants it predicted. A complete comparison between the Accuracy, PPV, NPV and MCC is shown in S3 Table.

Comparison with other models

To the best of our knowledge, ours are the only models trained on an ACMG/AMP gold standard dataset specifically created for *ATP7B* variants linked with Wilson's Disease. While other deep learning models based on ACMG/AMP guidelines such as RENOVO [21] and MLVar [22] exist, they are either not trained on manually classified variants/attributes, or do not follow a disease-specific approach. As each disease follows different genetic mechanisms, generalization for all is difficult to achieve by a single model. We have, however included scores generated by running RENOVO, as well as pre-determined scores obtained for 11 other

models including AlphaMissense, REVEL, SIFT, Polyphen2, Eigen-PC, LRT, MutationTaster, FATHMM, PROVEAN, MetaLR, and MutationAssessor [23–33] in S3 Table, and as S8 Fig. Our models were able to outperform the others over the combined metrics of Accuracy, PPV, NPV, and MCC.

Reclassification of VUS variants

We collected all *ATP7B* variants of unknown significance and conflicting or missing classification from the ClinVar [34] database as well as our in-house data and used the model to reclassify them. Out of 977 exonic variants, TabNet reclassified 736 variants as pathogenic and 241 as benign. XGBoost on the other hand reclassified 800 as pathogenic and 177 as benign. Overall, a 91.4% concordance in predictions (726 pathogenic and 167 benign variants) was observed between the two models. The complete list of these variants and their reclassification can be accessed in S4 Table.

Scatterplots of class probability vs the predicted class for each model across 251 exonic VUS variants that were a part of our validation dataset are shown in S9 and S10 Figs for TabNet and XGBoost respectively.

Patient data validation

Impacts of WT *ATP7B* protein variants in cellular copper excretion. The copper assay data for the *ATP7B* variants, S855F and C271X (positive control for impaired *ATP7B*) showed reduced copper levels in the media in comparison to the WT *ATP7B* (Fig 3). This implies that WT *ATP7B* promotes the excretion of excess copper in the media while mutants, S855F and C271X show impaired protein function.

We ran both our models on each of the variants: both models accurately identified the control C271X variant as pathogenic, and also classified S855F as pathogenic. Thus, both our models tested on functionally proven data provide accurate classifications of the variant.

Feature importance and computational efficiency

The feature importance of the top 20 features are depicted in S11 Fig. The larger the score, the higher the impact of the feature on the model. Both models had 10 features in common, relying on Loss of function (LoF) information, wherein a genetic lesion prevents the formation of a normal gene product thereby leading to disease. They also take into account the genomic position of the mutation (Start: nucleotide), which could dictate a pathogenic effect. Additionally they rely on global prevalence of variants (1000Genomes AF -ALL)—the number of high frequency disease causing variants is usually small, i.e. most pathogenic variants are rarely prevalent. The remaining features common to both models consist of pathogenicity scores obtained from 7 prediction tools (MetaSVM, MCAP, MutPred, SIF4G, REVEL, PolyPhen2 HDIV, and MutationTaster). Additional details of these features can be seen in S1 Table.

The XGBoost model additionally relies on the exonic function of the variant (Function), i.e. the nature of the effect the variant has (a stopgain/loss variant for example, would have a larger effect on the protein than a synonymous variant). It also takes into account the allele frequencies reported in the GnomAD database, which is a larger population dataset. Finally, it also considers conservation scores (Siphy 29way logOdds and MutationAssessor) that dictate how conserved a given site is among mammals, indicating a potentially important location, and thus a potentially more disruptive effect, as well as additional pathogenicity scores (DANN, MetaRNN, and BayesDel).

The TabNet model additionally considers variant prevalence in Gnomad (GnomadAF—Raw) and the Northeast African subset of the Greater Middle East populations (GME AF—

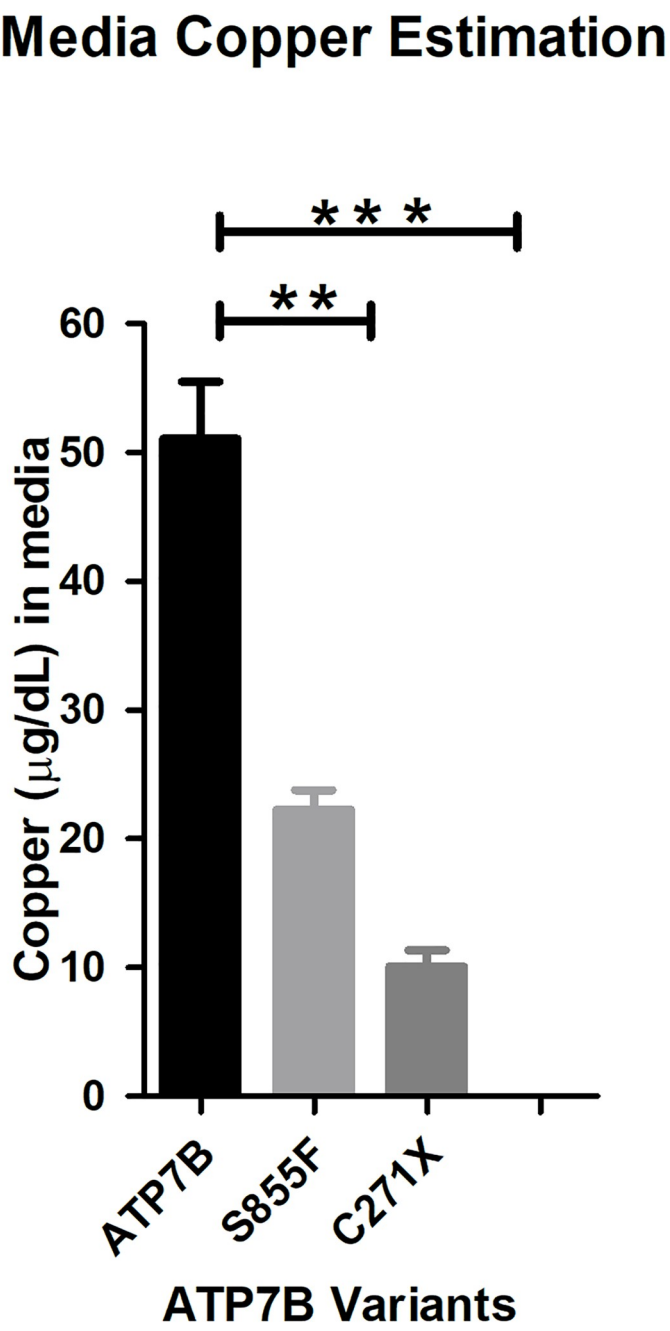

Fig 3. Copper exposure in *ATP7B* Knock-out HepG2 cells overexpressed with the plasmid containing wildtype and mutant *ATP7B* gene. The copper transport activity of *ATP7B* mutants S855F and C271X is significantly impaired in comparison to the wild-type *ATP7B*, where N = 3, ** denotes p value less than 0.01 and *** denotes p value less than 0.001.

https://doi.org/10.1371/journal.pone.0303787.g003

NEA), as well as pathogenicity and conservation scores (LRT, integreated_fitCons, PrimateAI, Eigen-PC-raw coding, and LIST-S2).

Thus both models take a well-rounded approach, and consider different aspects that determine variant pathogenicity, and are thus able to make reliable predictions. Further, the train

dataset labels have been determined through ACMG classification that take into account all aspects of relevant biological data including functional and segregational evidence. As such the models capture patterns among the attributes that lead to these classifications.

The complete time taken to process a VCF file into suitable input, and then train a model was plotted for each model separately, and are shown in S12 Fig.

Discussion and conclusion

In our work we have created two tools that can be used to classify variants of the *ATP7B* gene linked with Wilson's Disease. While tree-based XGBoost is one of the most reliable algorithms for tabular data, our study shows that TabNet, a deep learning model designed specifically for tabular data, slightly outperforms it in the classification of *ATP7B* variants. We have trained these models on a dataset classified through the application of ACMG guidelines, the gold standard in variant classification. Additionally, the data is a robust compilation of all publicly available variants of the gene exhaustively collected from literature and across 9 large databases. Thus the models were trained on accurately classified variants that capture all currently known types of exonic variants associated with Wilson's Disease, due to which we anticipate the models to be able to generalize to newly reported variants in the future. We have shown our models' accuracy through functional validation as well as comparison with other models. Finally, to address the large numbers of already reported variants of uncertain significance, we have collected and classified 977 exonic variants through both models; the predictions achieved a 91.4% concordance across 726 pathogenic and 167 benign variants. We have made these predictions openly available, along with their class probabilities to facilitate a better understanding of variant pathogenicity for clinicians and researchers.

Clinical diagnosis of Wilson's Disease is often challenging due to the heterogenous nature of symptoms it presents with. Genetic testing has thus been included in the diagnosis process as part of the Leipzig scoring system [23]. Additionally, testing can also rule out other genetic disorders such as some congenital disorders of glycosylation that mimic Wilson disease, but are not caused by *ATP7B* variants [35]. Since early diagnosis may prevent patients ever becoming symptomatic, infant and newborn screening, as well as family screening also become important. Accurate clinical interpretation of variants is therefore essential for diagnosis. Our models offer a means of applying learning of patterns based on classification by ACMG rapidly to a large number of variants, which otherwise is a time consuming and expertise-dependent process. Given the complex nature and varied mechanisms of genetic diseases, adopting a generalized approach to classifying causative variants is ill advised. We have shown this through the superior performance of our models over other general ACMG based models. To the best of our knowledge, no other models based on the ACMG classification of Wilson's disease variants currently exist.

We believe therefore, that our models can be utilized for the rapid classification of Wilson's Disease variants for better understanding of their pathogenicity in both research and clinical settings.

Limitations: Even though the WilsonGen database is an exhaustive compendium of currently known and classified variants, the number of classified exonic variants still remains small. ACMG classification of variants is also a time-consuming process, and thus a newer dataset may take time to make. We have thus been able to test model generalization on a dataset of 156 test variants. Additionally, the functional classification of ATP7B variants is still ongoing. Upon its completion, a clearer picture of which of the two models has performed better will be able to be obtained.

Supporting information

S1 Fig. Train and validation accuracies of the TabNet model across 187 epochs.
(TIF)

S2 Fig. Confusion matrix depicting the models' predictions on the 30% test data. Fig A represents TabNet while B represents XGBoost.
(TIF)

S3 Fig. The Precision-Recall curve of both the models.
(TIF)

S4 Fig. The model learning rate and loss of the TabNet model across187 epochs.
(TIF)

S5 Fig. Confusion matrix of predictions made on the ACMG-qualified independent validation dataset comprising of 156 variants. Fig A represents TabNet while B represents XGBoost.
(TIF)

S6 Fig. Scatterplot of class probability vs the actual ACMG class for the TabNet model across the validation set of 156 variants.
(TIF)

S7 Fig. Scatterplot of class probability vs the actual ACMG class for the XGBoost model across the validation set of 156 variants.
(TIF)

S8 Fig. Barplot comparing the accuracy, MCC, NPV and PPV of 13 models with TabNet and XGBoost. Abbreviations: MAssessor—MutationAssessor; MTaster—MutationTaster.
(TIF)

S9 Fig. Scatterplot of class probability vs the predicted class for the TabNet model across all VUS variants 251 exonic VUS variants that were a part of the validation dataset.
(TIF)

S10 Fig. Scatterplot of class probability vs the predicted class for the XGBoost model across all VUS variants 251 exonic VUS variants that were a part of the validation dataset.
(TIF)

S11 Fig. Plot depicting the feature importance of the top 15 features of each model. The x-axis for XGBoost plots F-score, while that of TabNet plots scores for each feature.
(TIF)

S12 Fig. Plot depicting the complete time taken to process a VCF file into suitable input, and then train a model was plotted for (A) TabNet and (B) XGBoost respectively.
(TIF)

S1 Table. A complete list of the 73 features used in training the model, along with the ACMG attribute they provide information about, along with their description, as well as their datatype.
(XLSX)

S2 Table. The classification report with the Precision, Recall and F1 scores for the TabNet model (A) and XGBoost model (B) respectively.
(XLSX)

S3 Table. Comparison of the performance of both models against 13 other models on the independent test dataset.
(XLSX)

S4 Table. A list of 977 exonic variants of uncertain significance reclassified by our models TabNet (A) and XGBoost (B). Variants highlighted in bold represent concordance between predictions from both algorithms. Table (C) describes the nucleotide and protein changes in HGVS nomenclature, and also describes each variant's exonic function.
(XLSX)

Acknowledgments

AV acknowledges a Senior Research Fellowship from ICMR. MK acknowledges a Senior Research Fellowship from CSIR.

Author Contributions

Conceptualization: Vinod Scaria, Binukumar B. K.

Data curation: Aastha Vatsyayan, Mukesh Kumar, Bhaskar Jyoti Saikia.

Formal analysis: Aastha Vatsyayan, Mukesh Kumar.

Funding acquisition: Vinod Scaria, Binukumar B. K.

Investigation: Vinod Scaria, Binukumar B. K.

Project administration: Vinod Scaria, Binukumar B. K.

Software: Aastha Vatsyayan.

Supervision: Vinod Scaria.

Validation: Mukesh Kumar, Bhaskar Jyoti Saikia.

Writing – review & editing: Mukesh Kumar, Vinod Scaria, Binukumar B. K.

References

1. Wang J, Tang L, Xu A, Zhang S, Jiang H, Pei P, et al. Identification of mutations in the ATP7B gene in 14 Wilson disease children: Case series. Medicine. 2021; 100: e25463. https://doi.org/10.1097/MD.0000000000025463 PMID: 33879678
2. Rodriguez-Castro KI, Hevia-Urrutia FJ, Sturniolo GC. Wilson's disease: A review of what we have learned. World J Hepatol. 2015; 7: 2859–2870. https://doi.org/10.4254/wjh.v7.i29.2859 PMID: 26692151
3. Gao J, Brackley S, Mann JP. The global prevalence of Wilson disease from next-generation sequencing data. Genet Med. 2019; 21: 1155–1163. https://doi.org/10.1038/s41436-018-0309-9 PMID: 30254379
4. Kim G-H, Yang JY, Park J-Y, Lee JJ, Kim JH, Yoo H-W. Estimation of Wilson's disease incidence and carrier frequency in the Korean population by screening ATP7B major mutations in newborn filter papers using the SYBR green intercalator method based on the amplification refractory mutation system. Genet Test. 2008; 12: 395–399. https://doi.org/10.1089/gte.2008.0016 PMID: 18652531
5. Roberts EA. Update on the Diagnosis and Management of Wilson Disease. Curr Gastroenterol Rep. 2018; 20: 56. https://doi.org/10.1007/s11894-018-0660-7 PMID: 30397835
6. Olivarez L, Caggana M, Pass KA, Ferguson P, Brewer GJ. Estimate of the frequency of Wilson's disease in the US Caucasian population: a mutation analysis approach. Ann Hum Genet. 2001; 65: 459–463. https://doi.org/10.1017/S0003480001008764 PMID: 11806854
7. Jang J-H, Lee T, Bang S, Kim Y-E, Cho E-H. Carrier frequency of Wilson's disease in the Korean population: a DNA-based approach. J Hum Genet. 2017; 62: 815–818. https://doi.org/10.1038/jhg.2017.49 PMID: 28515472

8. Collet C, Laplanche J-L, Page J, Morel H, Woimant F, Poujois A. High genetic carrier frequency of Wilson's disease in France: discrepancies with clinical prevalence. BMC Med Genet. 2018; 19: 143. https://doi.org/10.1186/s12881-018-0660-3 PMID: 30097039
9. 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: 405–424. https://doi.org/10.1038/gim.2015.30 PMID: 25741868
10. Kumar M, Gaharwar U, Paul S, Poojary M, Pandhare K, Scaria V, et al. WilsonGen a comprehensive clinically annotated genomic variant resource for Wilson's Disease. Sci Rep. 2020; 10: 1–6. https://doi.org/10.1038/s41598-020-66099-2 PMID: 32493955
11. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010; 38: e164. https://doi.org/10.1093/nar/gkq603 PMID: 20601685
12. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020; 581: 434–443. https://doi.org/10.1038/s41586-020-2308-7 PMID: 32461654
13. A global reference for human genetic variation. Nature. 2015; 526: 68–74. https://doi.org/10.1038/nature15393 PMID: 26432245
14. Scott EM, Halees A, Itan Y, Spencer EG, He Y, Azab MA, et al. Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery. Nat Genet. 2016; 48: 1071–1076. https://doi.org/10.1038/ng.3592 PMID: 27428751
15. Witten IH, Frank E, Hall MA. Data Mining: Practical Machine Learning Tools and Techniques. Elsevier; 2011. https://play.google.com/store/books/details?id=bDtLM8CODsQC.
16. Shwartz-Ziv R, Armon A. Tabular data: Deep learning is not all you need. Information Fusion. 2022. pp. 84–90. https://doi.org/10.1016/j.inffus.2021.11.011
17. Arik SO, Pfister T. TabNet: Attentive Interpretable Tabular Learning. 2019 [cited 25 Jul 2022].
18. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. 2016 [cited 25 Jul 2022].
19. Ketkar N, Moolayil J. Automatic Differentiation in Deep Learning. Deep Learning with Python. 2021. pp. 133–145.
20. Anaconda Documentation—Anaconda documentation. [cited 26 Jul 2022]. https://docs.anaconda.com/.
21. Favalli V, Tini G, Bonetti E, Vozza G, Guida A, Gandini S, et al. Machine learning-based reclassification of germline variants of unknown significance: The RENOVO algorithm. Am J Hum Genet. 2021; 108: 682–695. https://doi.org/10.1016/j.ajhg.2021.03.010 PMID: 33761318
22. 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: 1–12. https://doi.org/10.1038/s41598-022-06547-3 PMID: 35169226
23. 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 [cited 28 Feb 2024]. https://doi.org/10.1126/science.adg7492 PMID: 37733863
24. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet. 2016; 99: 877. https://doi.org/10.1016/j.ajhg.2016.08.016 PMID: 27666373
25. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003; 31: 3812. https://doi.org/10.1093/nar/gkg509 PMID: 12824425
26. Adzhubei I, Jordan DM, Sunyaev SR. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. Curr Protoc Hum Genet. 2013;0 7: Unit7.20. https://doi.org/10.1002/0471142905.hg0720s76 PMID: 23315928
27. Ionita-Laza I, Mccallum K, Xu B, Buxbaum J. A SPECTRAL APPROACH INTEGRATING FUNCTIONAL GENOMIC ANNOTATIONS FOR CODING AND NONCODING VARIANTS. Nat Genet. 2016; 48: 214. https://doi.org/10.1038/ng.3477 PMID: 26727659
28. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res. 2009; 19: 1553–1561. https://doi.org/10.1101/gr.092619.109 PMID: 19602639
29. Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods. 2010; 7: 575–576. https://doi.org/10.1038/nmeth0810-575 PMID: 20676075
30. 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. 2013; 34: 57. https://doi.org/10.1002/humu.22225 PMID: 23033316

31. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. PLoS One. 2012; 7: e46688. https://doi.org/10.1371/journal.pone.0046688 PMID: 23056405

32. 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. 2014; 24: 2125–2137. https://doi.org/10.1093/hmg/ddu733 PMID: 25552646

33. Reva B, Antipin Y, Sander C. Determinants of protein function revealed by combinatorial entropy optimization. Genome Biol. 2007; 8. https://doi.org/10.1186/gb-2007-8-11-r232 PMID: 17976239

34. 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. https://doi.org/10.1093/nar/gkt1113 PMID: 24234437

35. Espinós C, Ferenci P. Are the new genetic tools for diagnosis of Wilson disease helpful in clinical practice? JHEP Rep. 2020; 2: 100114. https://doi.org/10.1016/j.jhepr.2020.100114 PMID: 32613181