Clinical Research Article
Journal of Intelligent Systems: Theory and Applications 8(1) (2025) 35-46
DOI: 10.38016/jista.1501164

Modified Hard Voting Classifier Implementation on MEFV Gene Variants Increases in Silico Tool Performance: A Novel Approach for Small Sample Size

Tarık Alay[1*], İbrahim Demir[2], Murat Kirisci[3]

[1*] Ankara Etlik Integrated Healthcare Campus, Ankara, Türkiye

[2] Turkish Statistical Institute (TUIK), Ankara, Türkiye

[3] Department of Biostatistics and Medical Informatics, Istanbul University-Cerrahpaşa, Istanbul, Türkiye
mtarikalay@gmail.com, idemir@gmail.com, mkirisci@hotmail.com

Abstract
Objective: There are a limited number of pathogenic variants known in the MEFV gene. In silico tools fail to classify many MEFV gene variants. Therefore, it is essential to implement novel approaches. Our goal is to develop a new strategy to solve the even number classification problem while improving MEFV gene variant prediction accuracy using small datasets.
Material - methods: First, we determined the optimal number of computational tools for the model. We then applied eight distinct ML algorithms on the training dataset containing MEFV gene variants using the determined tools. We initiated the application of modified hard voting machine learning algorithms, using a training and validation dataset. Subsequently, we implemented a comparative analysis between the prediction results and existing algorithms and studies. Finally, we evaluated the gene and protein level ascertainment to identify hotspot regions.
Results: The ensemble classifier scored an average ROCAUC of 88%. The modified hard voting method correctly classified all known variants with 82% accuracy, outperforming both the soft voting (75%) and hard voting (70%) methods. The results showed that the prevalence of LP variants was approximately 2.5 times higher in domains compared to LB variants(χ2: 13.574, $p < 0.001$, OR: 2.509 [1.532-4.132]).
Conclusion: Considering the limited understanding of the clinical implications associated with MEFV gene mutations, employing a modified hard voting classifier approach may improve the classification accuracy of computational tools.

Keywords: Classification, FMF, Machine learning, MEFV, Voting Classifier

MEFY Gen Varyantlarında Modifiye Edilmiş Sert Oylama Sınıflandırıcısı Uygulaması, In-Silico Araç Performansını Artırıyor: Küçük Örneklem Boyutu İçin Yeni Bir Yaklaşım

Öz
Amaç: MEFV geninde bilinen sınırlı sayıda patojenik varyant bulunmaktadır. İn siliko araçlar, birçok MEFV gen varyantını sınıflandıramamaktadır. Bu nedenle, yeni yaklaşımların uygulanması gerekmektedir. Sert oylama sınıflandırıcıları ve sağlam doğrulama teknikleri sınıflandırma için kullanılabilir; ancak çift sayı sınıflandırması doğru bir şekilde yapılamamaktadır. Amacımız, hem çift sayı sınıflandırma sorununu çözmek hem de küçük veri setleri kullanarak MEFV gen varyantı tahmin doğruluğunu artırmak için yeni bir strateji geliştirmektir.
Yöntem: İlk olarak model için optimal sayıda hesaplama aracını belirledik. Daha sonra, belirlenen araçlar kullanılarak MEFV gen varyantlarını içeren eğitim veri setinde sekiz farklı makine öğrenme algoritması uygulandı. Eğitim ve doğrulama veri setinin kullanımıyla, modifiye edilmiş sert oylama makine öğrenme algoritmalarının uygulanmasına başlandı. Bundan sonra, tahmin sonuçları ile mevcut algoritmalar ve çalışmalar arasında karşılaştırmalı bir analiz gerçekleştirildi. Son olarak, gen ve protein düzeyinde değerlendirme yapılarak hotspot bölgeler belirlendi.
Bulgular: Topluluk sınıflandırıcısı, ortalama ROC AUC puanlarının %88 olduğunu gösterdi ve modifiye edilmiş sert oylama sınıflandırıcı yöntemi ile bilinen tüm varyantları %82 doğrulukla sınıflandırdı. Bu oran, hem yumuşak (%75) hem de sert oylama sınıflandırıcı (%70) yöntemlerinden daha yüksektir. Tüm varyantların kolektif değerlendirilmesi, LP varyantlarının, LB varyantlarına göre alanlarda yaklaşık 2,5 kat daha yaygın olduğunu ortaya koymuştur (χ2:13.574, $p < 0.001$, OR: 2.509 [1.532-4.132]).

* Corresponding Author.
E-mail: mtarikalay@gmail.com

Recieved : 14 June 2024
Revision : 22 Nov 2024
Accepted : 22 Jan 2025

Sonuç: MEFV gen mutasyonlarının klinik sonuçlarıyla ilgili bilgi yetersizliği göz önüne alındığında, modifiye edilmiş sert oylama sınıflandırıcı yaklaşımını kullanmak, hesaplama araçlarının sınıflandırma doğruluğunu artırmak için küçük örneklemlerde makul bir yöntem olabilir.
Anahtar Kelimeler: Sınıflandırma, Ailevi Akdeniz Ateşi, Makine Öğrenmesi, *MEFV*, Oylama Sınıflandırıcısı

1. Introduction

The widespread utilization of next-generation sequencing technology enhances the probability of diagnosing familial Mediterranean fever (FMF), ascertains the carrier rates within the population, and forecasts the likelihood of disease recurrence. Although the widespread utilization of Next-Generation Sequencing (NGS) assays has led to the discovery of a multitude of novel variants within the MEFV gene(Kırnaz, Gezgin and Berdeli, 2022)[1], The International Study Group on Systemic Autoinflammatory Disorders (INSAID) consensus criteria found that the clinical outcomes of more than half of the MEFV gene variants are categorized as variation of unknown significance fort the American College of Medical Genetics (ACMG)(Van Gijn *et al.*, 2018) [2].

Physicians and patients face difficulties in comprehending and interpreting the clinical implications of variants of uncertain significance (VOUS). In order to ascertain the clinical implications of the VOUS variant, it is necessary to conduct well-executed functional and hereditary investigations. However, these studies are associated with substantial costs and time requirements. Consequently, there is a need for innovative approaches that are both rapid and cost-effective, while also posing minimal risk, to predict the consequences of MEFV variants(Richards *et al.*, 2015; Nykamp *et al.*, 2017) [3, 4]. The utilization of existing variant prediction tools was considered as the second option. Nevertheless, there is a divergence of viewpoints regarding the selection and utilization of protein prediction methods and meta-predictors for the purpose of clinical variant evaluation, as highlighted by Richards et al. (Richards *et al.*, 2015)[3]. The ACMG and Clingen organizations have recommended conducting extensive evaluations at the gene level(Pyeritz and for the Professional Practice and Guidelines Committee, 2012; Stewart *et al.*, 2018; Harrison, Biesecker and Rehm, 2019; Burdon *et al.*, 2022; Lai *et al.*, 2022)5–9]. Although despite these efforts, it is still insufficient to accurately predict the clinical implications of most genes, including MEFV.

Our research endeavours focused on the exploration of a novel approach that incorporates an optimal selection of tools and machine learning algorithms, aiming to achieve a level of accuracy that is close to perfection. The accuracy of predicting outcomes is dependent on the training data exhibiting high levels of responsiveness. Therefore, the implementation of novel machine learning selection methods is expected to mitigate uncertainties. Nevertheless, numerous machine learning algorithms are currently employed in various amino acid prediction scores, meta scores, and ensemble algorithms. However, conventional machine learning (ML) algorithms are developed by choosing the classification method that yields the highest level of accuracy. This process fails to adequately acknowledge the success achieved by other machine learning algorithms. Many ML algorithms work well in large datasets(Song *et al.*, 2021). However, some datasets contain many uncertainties, making it impossible to achieve larger sample sizes(Accetturo, Bartolomeo and Stella, 2020; Alay, 2024). Therefore, in these situations, it is imperative to develop novel methodologies. Hard and soft voting classifiers are employed to enhance the performance of in silico tools and to evaluate the contribution of multiple scores to the classification process. However, hard voting classifiers perform binary classification (1 or 0), making accurate assignments in cases involving an even number of classifiers challenging. Many previous studies have reported difficulties in achieving consensus with an even number of algorithms(Awe *et al.*, 2024). To address this specific limitation, it may be beneficial to develop a method that incorporates only the most effective algorithms into the prediction process.In this study, we propose and evaluate a method called the "modified hard voting classifier" designed to overcome this issue.

This study aims to present a novel methodology for improving the accuracy of *MEFV* gene variant classification by utilizing optimal amino acid prediction scores and machine-learning algorithms. Our objective is to establish a more precise categorization of *MEFV* variants while minimizing uncertainty through the development of a new voting classifier. The findings of this study will provide valuable insights for clinicians in interpreting the clinical significance of variants with ambiguous effects on health outcomes and contribute to the development of gene-specific interpretation guidelines.

2. Material-Methods

2.1. Machine Learning Analyses

Libraries

Python were utilized for machine learning analysis step. The following libraries were utilized: sklearn for machine learning analysis, seaborn and matplotlib for data visualization, statsmodel for statistical models, and pandas and numpy for data manipulation. All versions of libraries were compatible with Python 3.7.1. Evaluation.

2.2. Data Retrieval Process

We obtained 389 MEFV variants from the Infevers database (https://infevers.umai-montpellier. fr/web, last access date:05/04/2022), focusing solely on single

nucleotide variants within the coding region such as missense and silent variants. Variants such as frameshift/inframe deletions, termination gain, termination loss, insertions, , and indels were omitted for analysis. In line with Clingen and ACMG guidelines, only clinically validated SNV predictors endorsed or evaluated by the Clingen group or ACMG guidelines were considered(Richards *et al.*, 2015; Ioannidis *et al.*, 2016; Tian *et al.*, 2019; Savige *et al.*, 2021; Pejaver *et al.*, 2022; Cheng *et al.*, 2023)

2.3.Feature Determination

During the process of selecting in-silico tools, we evaluated a number of conditions. First, we chose *in-silico* tools because they were up-to-date, validated, and recommended by ACMG guidelines(Richards *et al.*, 2015) and Clingen Group (Savige *et al.*, 2021; Waring *et al.*, 2021; Pejaver *et al.*, 2022; Wilcox *et al.*, 2022). Second, we meticulously determined that missing values for relevant variant scores in the entire dataset should not be included(Palanivinayagam and Damaševičius, 2023). Third, we compared all scores multicollinearity by using Spearman correlation. According to these rules, four in silico tools (Revel,MetaLR,SIFT, FATHM) were detected compatible with our algorithms. Other details of the selection *in silico tool* process are indicated in Supplementary File 1.

2.3. Feature Engineering

Data Preprocessing

For encoding dummy variables, the "Label Encoder" and "Ordinal Encoder" methods of sklearn.preprocessing are utilized. The "standard scaler" method was implemented for data standardization. The standard scaler method implemented after dataset split into training validation and prediction.

$\frac{X-u}{z}$ **(1)** Standardization(Z-score normalization)

Checking for Normality, Data Transformation and Dimension Reduction

We examined the distribution patterns for four distinct scoring metrics and determined that three of them were right-skewed, while the remaining one was left-skewed. In response to these findings, we applied square root and logarithmic transformations to normalize our dataset. Given the non-normal distribution of all four scores, we employed the Kruskal-Wallis H test as the appropriate non-parametric statistical method.

After conducting a thorough investigation, we found a total of 266 distinct MEFV gene mutations in our dataset. The breakdown can be outlined as follows: The recorded values are as follows: The distribution of the classifications of the variations is as follows: Benign (B): 3, Likely Benign (LB): 46, Likely Pathogenic (LP): 44, Pathogenic (P): 5, Variations of Unknown Significance (VOUS): 110, Not Categorized (NC): 26, Unsolved (US): 32. Given the diverse attributes of this dataset and the challenges associated with the seven-tier classification system, we recognized the need to decrease the number of dimensions to achieve a fairer and more understandable analysis. Several prior research employed the same methodology. (Accetturo *et al.*, 2020; Accetturo, Bartolomeo and Stella, 2020; Mighton *et al.*, 2022) Figure 1 provides a visual depiction and comparative examination of the seven-tier and three-tier classification systems. By employing dimensionality reduction techniques, we have circumvented the "curse of dimensionality," thereby defining boundaries that facilitate more accurate discrimination between damaging and benign genetic variants.

Upon conducting an extensive review, we identified 98 clinically recognized variants, evenly split between 49 likely benign and 49 likely pathogenic. These variants were selected to form a balanced training dataset. The dataset was subsequently partitioned, allocating 80% (n=78) for training purposes and 20% (n=20) for validation. Utilizing this set of clinically

Figure 1. MEFV variants distributions according to seven-tier and three-tier categories. For this step, we verified whether REVEL, SIFT, MetaLR, and FATHMM scores for the classes "LB" vs. "B" and "LP" vs. "P" did not show a statistically significant difference. In no cases, the medians of the two benign and two pathogenic classification groups demonstrated statistically significant differences (Kruskal-Wallis test not significant for non-parametric ANOVA of "LP" vs. "LP/P vs. "P" and "B" vs. "LB"). Therefore, we merged into LB and B as LB, LP and P as LP, and NC, VOUS, and US as VOUS. a1,b1) The frequency of variants according to infevers seven-tier classification system and our three-tier classification system,respectively. a2,a3,a4,a5) The box and plot distribution of infevers seven-tier classification according to Revel, MetaLR, SIFT, and FATHMM classification systems. b2,b3,b4,b5) The box and plot distribution of new three-tier classification according to Revel, MetaLR, SIFT, and FATHMM classification systems c1,d1) Comparison of variants according to exonic placements betwen seven-tier Infevers Classification and three-tier new classification system. Most of the pathogenic variants were placed in exon 10. c2,c3,c4,c5) Variant distribution by cDNA position according to Infevers seven-tier classification system. No certain pattern of clustering detected d2,d3,d4,d5) Variant distribution by cDNA position according to new three-tier classification system. Higher than 0.9 Revel scores most likely associated with variant pathogenicity similar to Clingen PP3

classification evaluation. No clear distinguished threshold is evident for other scores e1,f1) Variant distribution according to pyrin protein domains. PF02758: PAAD/DAPIN/pyrin domain, PF00643: Domain b-Box Zinc Finger domain, PF13765:SPRY-associated domain, PF00622: SPRY domain Most of the pathogenic variants placed in SPRY domain of pyrin protein. e2,e3,e4,e5) Variant distribution by aminoacid position according to infevers seven-tier classification system. f2,f3,f4,f5) Variant distribution by aminoacid position according to new three-tier classification system

Corroborated variants, our aim was to ascertain the optimal number of features necessary for reliable predictions. A review of existing literature, coupled with sample size determinations, revealed that a quartet of in-silico tools yielded the most favorable performance (Ogundimu, Altman and Collins, 2016; Riley *et al.*, 2019; Accetturo *et al.*, 2020; Acharjee *et al.*, 2020; Luan *et al.*, 2020).

2.4. Feature Selection

2.4.1. Selection of Machine learning Methods

We utilized seven machine learning techniques—K-nearest neighbor (KNN), Decision Tree(DT), Random Forest (RF), Multilayer perceptron Logistic regression (LR), Linear Support Vector Machine (SVM-linear), and Radial basis function Support Vector Machine (SVM-RBF)—to analyze four scores (SIFT, FATHMM, Revel, and MetaLR).

2.4.2. Dataset Evaluation

We trained RF, DT, KNN, LR, LSVM, KSVM, and PSVM on four scores (REVEL, MetaLR, SIFT, FATHMM). We did k-fold crossvalidation, leave one out of crossvalidation, leave p out of crossvalidation, and validation dataset techniques for a model validation and generalizability techniques. As compatible with our dataset nature k-fold cross-validation put forward best results with 10 values. As our training dataset was balanced, so we determined our threshold value according to the accuracy score. We used other paramaters such as precision, recall and F1 metrics for dataset evaluation explained at Supplementary File S2.

2.4.3. Modified Hard Voting Classifier

Problem

A Hard Voting Classifier cannot make an assignment in the case of a tie. In such instances, weighting is necessary; however, applying weights requires prior assumptions about these characteristics. This approach neglects the individual performance metrics of the algorithms. A new classification method that considers performance metrics for binary classification could resolve this issue for the Hard Voting Classifier.

Formulas

Given a set of n algorithms $\{A_1, A_2, \ldots, A_n\}$, where n is an even number and $\frac{n}{2}$ is an odd number, we aim to select machine learning algorithms with the highest ROCAUC scores and use hard voting to combine their predictions. Let the ROCAUC scores of algorithms be $\{ROC_{A1}, ROC_{A2}, \ldots, ROC_{An}\}$.

1. **Select the algorithms with ROC AUC scores greater than 0.80:**
   Successful_algorithms = $\{ A_i \mid ROC_{A_i} > 0.80 \}$
2. **Iterative Reduction to an Odd Number**
   While the number of successful algorithms is even and greater than 1, reduce it by half: While $|Successful_{algorithms}| \% 2 = 0$ and $|Successful_{algorithms}| > 1$:
   Successful_algorithms = $\{ Ai \mid Accuracy_{A_i} \text{ in top half of } Accuracy \text{ scores of } Successful\_algorithms \}$
3. **Final Set of Algorithms**
   After the iterative reduction, let $\{A_{f1}, A_{f2}, \ldots, A_{fm}\}$ be the final set of algorithms, where m is an odd number.
4. **Hard Voting Classifier**
   Combine the predictions of the final set of algorithms using a hard voting mechanism: $\hat{y} = argmax_k \sum_{j=1}^{m} \chi (C_{A_{fj}}(x) = k)$
   $C_{A_{fj}}(x)$ is the prediction of algorithm for A_{fj} instance x, and χ is the indicator function that equals 1 if the condition is true and 0 otherwise.

Application

We conducted assessments of machine learning algorithms with a focus on those that exceeded a pre-established accuracy threshold. Our analysis techniques were built on ensemble models, specifically using a voting prediction approach. This method did not assign weighted scores for predictive accuracy; instead, our classification system was binary, labeling outcomes as either "classified" or "not classified." For instance, should all three machine learning algorithms concur in identifying variant "X" as LP, it would receive a score of “3” and be categorized accordingly as LP. Conversely, if only two algorithms determined variant "Y" to be LB, it would garner a score of “2” and be categorized as LB. Thus, our method operates under stringent criteria without utilizing weighted scoring, leading us to describe it as a “modified hard voting classifier.”

Consequently, we predicted our variants similar to the hard voting classifier algorithm. However, four approaches were different from the hard voting classifier: (1) Voting classifier did not solve even numbers classification problem. (2)We did not only implemented hard voting classifier on not only training and validation dataset, but also prediction scores. (3) Different from classic hard voting classifier we did not calculate all scores, we only included voting a showed outstanding area under curve scores which was accepted as higher than 80%. (4) We selected each algorithms

best paramaters not only combination of best paramaters of scores [Figure 2].

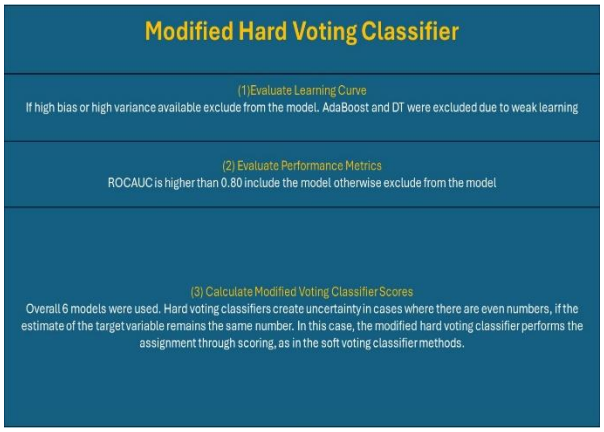

Figure 2. Establishing a modified Hard Voting Classifier

2.5. Functional and Clinical Level Evaluation

We evaluated each variant in two categories for functional-level ascertainment: gene-level and protein-level. While we established gene-level evaluation based on exonic position, we implemented protein-level evaluation by comparison of pyrin protein domain distributions. We evaluated *MEFV* domains initiation and termination location according to protein databank (https://www.rcsb.org/), Ensemble, Prosite (https://www.expasy.org/resources/prosite), conserved domain databases (https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml), InterPro (https://www.ebi.ac.uk/interpro/), and existing literature(Grandemange *et al.,* 2011). *MEFV* (NM_000243.3)

transcripts were based on variants distribution on pyrin protein.

2.6. Sample Size Calculation

Before implementing machine learning analysis, we had to conduct sample size calculations. Our sample size calculation was implemented on the basis of study of Accetturo et al. (Accetturo *et al.*, 2020). Furthermore, the metapredictors and aminoacid prediction tools that we implemented in the study have already been trained on a larger dataset(Waring *et al.*, 2021; Pejaver *et al.*, 2022; Sallah *et al.*, 2022). Considering the number features and sample size it is sufficient to implement machine learning methods(Ogundimu, Altman and Collins, 2016; Riley *et al.*, 2019; Luan *et al.*, 2020; Rajput, Wang and Chen, 2023).

2.7. Statistical Analysis

Our statistical analyses were performed utilizing Python version 3.7.1 alongside SPSS version 25.0 for Windows (IBM, Chicago, IL). We established a 95% confidence interval for the entirety of our statistical tests. The significance levels, denoted as alpha (α) and beta (β), were set at 0.05 and 0.20, respectively. A p-value threshold was determined to be 0.05, with values falling below this cutoff being considered statistically significant.

To evaluate the distribution of both discrete and continuous numerical variables, normality was probed using a suite of graphical and analytical techniques. Conformity with normal distribution assumptions allowed the use of means and standard deviations; in their absence, medians and interquartile ranges were employed. Categorical variables, either nominal or ordinal, were quantified and expressed as frequencies and percentages, with ordinal variables arranged according to their inherent hierarchy.

Graphical methods such as Q-Q plots, detrended plots, boxplots, histograms, and stem-and-leaf plots, alongside the analytical Kolmogorov-Smirnov test, were utilized to assess the normality of the data. The range for skewness and kurtosis was considered acceptable between -1 and +1, while skewness and kurtosis indices – calculated by dividing the respective values by their standard errors – were deemed to reflect normality when falling within the -2 to +2 range.

For variables that adhered to normal distribution, we applied the Analysis of Variance test. This was followed by post hoc analysis using the Tukey test to identify significant pairwise differences. In the case of non-normally distributed data, the non-parametric Kruskal-Wallis H-test was administered, succeeded by the Dunn-Bonferroni test for post hoc comparisons.

3. Results

3.1. Evaluation of Training dataset

It is highly recommended that, when developing novel algorithms, one should not only concentrate on novelty but also identify a good feature dataset(Khalid and Sezerman, 2018). Therefore, we examined our training dataset and determined the threshold of 80% ROCAUC required for machine learning algorithms to succeed. All algorithms exceeded the threshold value. However, Adaboost and DT were excluded from the model due to their overfitting[Supplementary Figure 3 and 4]. Revel score is detected as the most important feature when classifying datasets according to the most accurate classifier algorithm, RF [Figure 3]

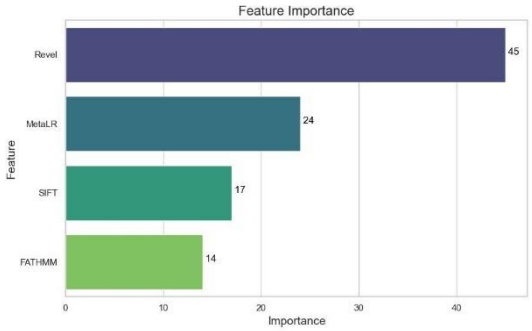

Figure 3. Feature Importance Metrics according to Random Forest Classifier. Revel is the most important feature which is contributed roughly about 50% to classifier.

3.2. Validation and hyperparamater tuning

Overall 98 known variants analyzed under two dataset: training dataset (n=78) and validation dataset (n=20). All. After implementing the machine learning algorithm, we conducted hyperparameter tuning for our accurate machine learning classifier algorithms [Table 1]. The learning curve demonstrates low bias and variance, even when trained on a small dataset, indicating robust and reliable model performance [Figure 4]. K-fold Crossvalidation (CV) and nested CV methods were used for validation methods which were more robust to sample size(Vabalas *et al.*, 2019; Larracy, Phinyomark and Scheme, 2021; Dalmaijer, Nord and Astle, 2022).

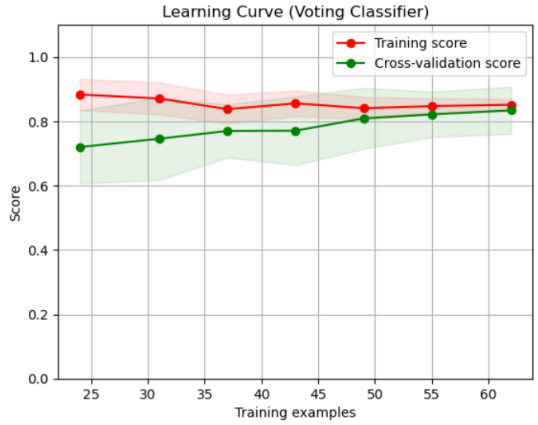

Figure 4. Modified Hard Voting Classifier Learning Curve. Both the training and validation curves were generated from a small dataset, exhibiting relatively low bias and variance, thus indicating a robust and reliable model performance.

Table 1. Cross validation results in validation dataset(n=78 for training dataset, and n=20 for validation dataset)

ML Methods	Precision	Recall	Accuracy	ROC AUC
LR	0.79	0.75	0.76	**0.9**
SVM-RBF	**0.81**	0.77	0.77	0.89
SVM-Linear	0.77	**0.86**	**0.81**	**0.9**
Gaussian NB	**0.89**	0.75	**0.81**	0.89
kNN	**0.82**	0.77	0.79	0.85
RF	**0.86**	**0.79**	**0.82**	**0.91**

Stratified K-fold CV implemented (The best results obtained in 10-fold CV. The 10-fold CV results were represented above). The results were checked with nested cv which is more robust to sample size. The similar results were obtained. The best parameters obtained for each classifier are as follows: **RF classifier**: - 'bootstrap': False - 'max_depth': None - 'min_samples_leaf': 3 - 'min_samples_split': 10 - 'n_estimators': 10 max_features=3 **KNN classifier**: - 'algorithm': 'auto' - 'n_neighbors': 3 - 'p': 2 - 'weights': 'uniform'. **DT classifier:** - 'criterion': 'entropy' - 'max_depth': None - 'min_samples_leaf': 2 - 'min_samples_split': 2 **NB Classifier** n_jobs=-1, cv=5, verbose=5, var_smoothing= 1e-6 , **LR Classifier** penalty=L2, C:1000, **SVM Classifier** 'C': 1000, 'gamma': 0.01, 'kernel': 'rbf' **SVM-linear Classifier** {'C': 1000, 'break_ties': False, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape': 'ovr', 'degree': 3, 'gamma': 0.01, 'kernel': 'rbf', 'max_iter': -1, 'probability': False, 'random_state': None, 'shrinking': True, 'tol': 0.001, 'verbose': False} **SVM-RBF classifier** {'C': 1000, 'break_ties': False, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape': 'ovr', 'degree': 3, 'gamma': 0.01, 'kernel': 'rbf', 'max_iter': -1, 'probability': False, 'random_state': None, 'shrinking': True, 'tol': 0.001, 'verbose': False}. For hyperparameter optimization, the GridSearchCV algorithm was implemented with a 10-fold CV. However, due to the overfitting observed in DT and AdaBoost Classifier, both of the algorithms were excluded from the analysis.

3.3. Comparison of the Training Dataset Results with Existing Literature

The next step we compared our results with existing literature and scores we used in our dataset. According to this comparison, modified hard voting classifier has most accurate classifying known variants [Figure 5]. The mean ROCAUC of six remaining ML methods was detected as 88% in both training and validation dataset. Interestingly, modified hard voting classifier classified more than 82% of known variants correctly in overall (training and validation) dataset. In the literature, the second most accurate classifier was Linear Discriminant Analysis conducted by Accetturo et al. classified variants with 75 % accuracy(Accetturo *et al.*, 2020). Most of the predictors classified LB variants with higher ROCUAC scores than 80%; however, LP classification showed a wide range of variety in accuracy scores between 2% - 62.5(Ioannidis *et al.*, 2016; Liu *et al.*, 2016; Knecht *et al.*, 2017; Tian *et al.*, 2019; Accetturo *et al.*, 2020).

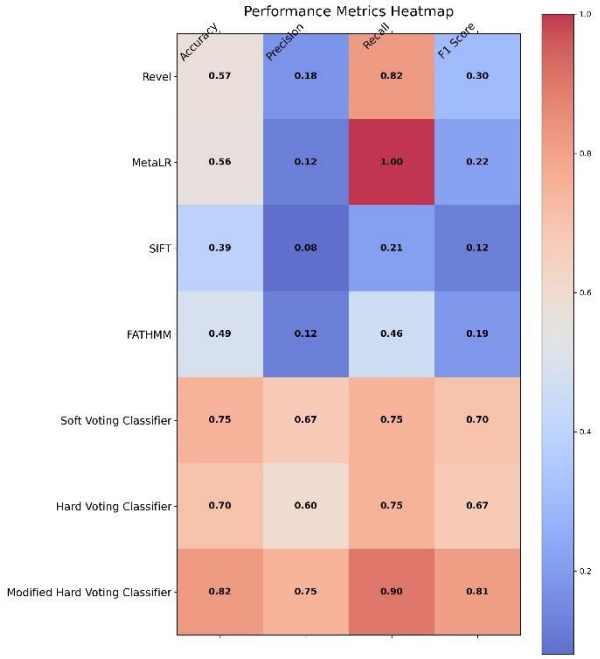

Figure 5. Comparison of modified hard voting classifier with existing algorithms by evaluating their success in classifying known variants. This figure illustrates the improved classification metrics achieved by a modified hard voting classifier for the prediction of MEFV gene variants. Traditional in silico predictors have struggled to distinguish MEFV gene variants, often performing at levels comparable to random chance. The modified hard voting classifier, however, demonstrates enhanced accuracy, sensitivity, and specificity, showcasing its superior discriminatory power in the analysis of MEFV gene variants. Additionally, this classifier has improved the classification performance of the existing hard voting classifier. As a result, it has outperformed the soft voting classifier. The modified hard voting classifier, especially for small sample sizes, can be combined with well-tuned k-fold cross-validation or nested CV methods, which are not significantly affected by the sample size.

3.4. Prediction Outcomes and Evaluation of Machine Learning Algorithms on VOUS variants

After the voting classification of training (n=78) and validation (n=20) dataset, overall 94 out of 98 (95.91%) variants were classified accurately in our dataset. The same prediction implemented for VOUS variants. Overall, we found 85 LP variants and 83 LB variants. As a result, we discovered 134 LP variants and 132 LB variants in the overall dataset. New distribution of all *MEFV* gene variants indicated in Figure 6.

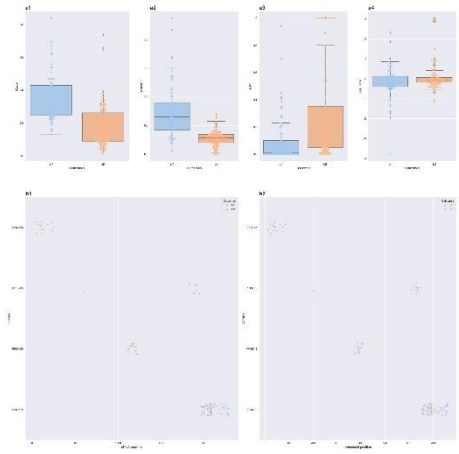

Figure 6. Visualization of Prediction Algorithm Results. a) While, Revel, MetaLR, and SIFT algorithms contributed statistically significant effect on model ($p<0.05$), FATHMM algorithm plays a supporting role on it. b1) Domain distribution of MEFV gene variants prediction results according to cDNA position b2) Domain distribution of MEFV gene variants prediction results according to aminoacid position. While most of the pathogenic variants distributed into PF00622 domain (2.595[1.525-4.425], $p<0.001$) , most of the benign variants distributed into PF00643 domain. PF02758: PAAD/DAPIN/pyrin domain, PF00643: Domain b-Box Zinc Finger domain, PF13765:SPRY-associated domain, PF00622: SPRY domain Most of the pathogenic variants placed in SPRY domain of pyrin protein.

3.5. Functional Evaluation

3.5.1. Gene-level (Exonic) Ascertainment

In our initial assessment of variants of uncertain significance, we ascertained that exon 10 harbored 37.6% (32/85) of variants predicted as likely pathogenic, whereas exon 2 contained 41.0% (34/83) of those predicted to be likely benign. Through our prediction methodology, it was concluded that 61.5% (32/52) of exon 10 variants and 58.6% (34/58) of exon 2 variants were classified as LP and LB, respectively. A disproportionate distribution was observed, with exons 7, 9, and 10 presenting a greater prevalence of LP variants in contrast to the preponderance of LB variants in other exons. In particular in exon 10, 42.5% (57/134) of the variants were categorized as likely pathogenic (LP), and in exon 2, 43.2% (57/132) were classified as likely benign (LB). Statistical analysis demonstrated a significant discrepancy in the distribution between LP and LB variants in these exons. As a result of our gene-level analyses revealed that exon 10 variants were 2.6 times more prone to be classified as LP than LB ($\chi2$: 12.858, $p < 0.001$, odds ratio [OR]: 2.629; 95% CI: 1.539-4.493). In contrast, exon 2 variants had a higher likelihood of being labeled as LB compared to LP ($\chi2$: 12.693, $p < 0.001$, OR: 2.595; 95% CI: 1.532-4.132). Afterwards, we combined our prediction outcomes with the training datasets (LB and LP) and assessed them according to exonic positions [Table 2].

Table 2. Distribution of all variants by exons and variant prediction outcomes

Exons	Variant prediction outcomes		p values	95% Interval		
	LB (n=132)	LP (n=134)		Odds ratios	Lower	Upper
1	4	11	0.118[a]	2.862	0.887	9.228
2	57	30	**<0.001**[b]	0.380	0.223	0.646
3	20	14	0.334[a]	0.653	0.315	1.356
4	3	2	0.683[c]	0.652	0.107	3.963
5	14	10	0.463[a]	0.680	0.291	1.590
6	-	-	*	*	*	*
7	-	4	0.122[c]	*	*	*
8	1	4	0.370[c]	4.031	0.445	36.549
9	4	2	0.445[c]	0.485	0.087	2.693
10	29	57	**<0.01**[b]	2.629	1.539	4.493

a Chi-square(Yates correction), b. Pearson chi-square test, c. Fisher Exact test,*not calculated. Evaluation of each exon is based on the LP to LB ratio.

3.5.2. Protein-Level (Domain-based) Evaluation

Within the domains, we properly identified 47% (40/85) predicted LP variants and 28.92% (24/83) predicted LB variants. After assessing the anticipated variations (n = 168), it was discovered that variants located within the domain were 2.766 times more likely to be classified as LP compared to LB (2:10.566, p:0.002, OR: 2.766 [1.462-5.233]). Subsequently, we combined the training dataset with the predicted VOUS variants. After collectively evaluating all variants, we found that LP variants were approximately 2.5 times more common in domains compared to LB variants (χ2:13.574, $p < 0.001$, OR: 2.509 [1.532-4.132]). On the other hand, B30.2 domain variants had a 2.5-fold higher likelihood of being LP compared to LB. This difference was statistically significant (χ2:12.693, $p < 0.001$, OR: 2.595 [1.532-4.132]). Nevertheless, the likelihood of variants that were not found in any domains being LB was 2.6 times higher compared to LP (χ2:14.508, $p < 0.001$, OR: 0.386 [0.235–0.633]). Upon identifying this statistically significant disparity, we assessed all variations within their respective domains [Table 3].

Table 3. Distribution of all variants by domains and variant prediction outcomes

Domains	Variant prediction outcomes		p values	95% Interval		
	LB (n=132)	LP (n=134)		Odds ratios	Lower	Upper
PYD	4	11	0.118[a]	2.862	0.887	9.228
bZIP	6	1	0.065[b]	0.158	0.019	1.330
B	4	8	0.390[a]	2.032	0.597	6.918
CC	4	2	0.445[b]	0.485	0.087	2.693
B30.2	31	58	**<0.001**[c]	2.595	1.525	4.425
Not identified	83	54	**<0.001**[c]	0.386	0.235	0.633

a Chi-square (Yates correction), b Fisher Exact test c. Chi-square test

4. Discussion

Many novel ML algorithms are designed to predict outcomes for larger datasets. However, few strategies are available for small datasets(Liu *et al.*, 2013; Vabalas *et al.*, 2019; Albaradei *et al.*, 2021; El-Sofany, Bouallegue and El-Latif, 2024). In this context, the modified hard voting classifier demonstrates superior performance, surpassing traditional hard voting and soft voting methods while effectively addressing challenges such as odd-number classification, which refers to scenarios where standard voting methods struggle to

make definitive decisions in cases with an uneven distribution of votes. By optimizing predictions, this approach enhances the accuracy of in silico tools and offers a reliable solution for analyses involving limited sample sizes.

New applications and Implementation Steps

This study includes a number of enhanced methods and new technologies. We base our new approach on a three-fold framework. The first step involves using big data analysis and comparing the datasets with existing algorithms and previous research findings. The second and the third steps include functional and protein-level evaluations, respectively.

Evaluation of Results

Initially, we applied seven machine learning algorithms on the training set, specifically the LP and LB variants. For the prediction of VOUS variants, we selected three out of the six machine learning techniques that had a minimum ROCUAC of 80%. We obtained 88% mean ROCAUC results for all 6 algorithms: LR, SVM-RBF, SVM-linear, Gaussian NB, KNN, RF. According to our sample, our voting classifier model correctly classified LB and LP variants. Subsequently, we assessed our training dataset by comparing it to established variant prediction tools and previous research. Based on the comparison results, the modified hard voting classifier method demonstrated superior performance in classifying MEFV variants compared to existing *in silico* algorithms and previous studies(Accetturo *et al.*, 2020). In the second and third steps, we conducted a comprehensive functional level analysis, evaluating all variants from both gene-level and protein-level perspectives. Our analysis at the functional level revealed that the SPRY domain(Papin *et al.*, 2007), which corresponds to exon 10 (Dundar *et al.*, 2022)and accounts for a significant portion of predicted damaging MEFV gene variants, exhibited a statistically significant increase in LP variants in non-evolutionarily conserved regions. However, this increase was nearly equivalent to that observed in other evolutionarily conserved regions, and the difference was not statistically significant when compared to these conserved regions.

Modified hard voting classifier

The modified hard voting classifier introduces several novelties in the literature. First of all, the modified hard voting classifier approach incorporates an optimal quantity of protein prediction tools(Ng and Henikoff, 2003) or meta-predictors(Ioannidis *et al.*, 2016). Additionally, this method assesses the influence of all effective machine-learning techniques. To the best of our knowledge, we have made the initial modification to a hard voting classifier for the purpose of variant classification and two distinct classification methods. Rather than relying on the traditional hard voting classifier, which explicitly votes on a single target variable, our approach utilizes both LB and LP variations, establishing a precise threshold for decision-making. Models that exhibited overfitting or underfitting were systematically eliminated, and the voting process was repeated until an optimized model was identified for predicting classification outcomes. Each model was evaluated with its own optimal parameters, ensuring rigorous performance testing. The modified hard voting classifier incorporates voting mechanisms to provide a rigorous classification procedure. Our high training data accuracy score stems from an optimum number of tools(Megantara and Ahmad, 2021; Hu *et al.*, 2024). In contrast to the first study on *MEFV* gene unknown variant prediction conducted by Accetturo et al.(Accetturo *et al.*, 2020), and existing tools , our prediction was derived from an ensemble method rather than relying on the most effective sole machine learning algorithm.

Literature review

The existing study provides a significant contribution to the literature by offering innovative solutions to three issues that previous *in-silico* tools have failed to address. The first issue is that current methods fail to successfully classify MEFV gene variants using numerous variant prediction algorithms (Accetturo *et al.*, 2020). Therefore, it is difficult to interpret variants according to current in silico tools (Ioannidis *et al.*, 2016). However, the modified hard voting classifier does not rely solely on a single *in-silico* tool or one ML method. The selection criteria of ML methods and *in-silico* tools are based on strict criteria, and only include most accurate methods or best features. Second, a significant issue is that during the variant classification process, many predictors correctly classify benign variants; however, many tools often fail to detect pathogenic variants accurately at the desired level (Adzhubei, Jordan and Sunyaev, 2013; Knecht *et al.*, 2017; Fortuno *et al.*, 2018; Pejaver *et al.*, 2022; Wilcox *et al.*, 2022). The comparative analysis revealed that our innovative methodology, the modified hard voting classifier, outperformed current in silico algorithms in classifying MEFV variants. This outcome arises from the modified hard voting classifier, which depends on a consensus of multiple machine learning techniques. Third, significant novel tools present better results day after day; unfortunately, still many variants remain unresolved. Even the newly developed in silico tool, Alphamissense, cannot classify 20% of all gene variants (Cheng *et al.*, 2023). The modified hard voting classifier effectively resolves uncertainties in variant interpretation.

Limitations of the study

Although this method produces very high classification rates, it has some drawbacks when applied to our dataset. First, identifying the optimal classifiers from among hundreds of in silico tools remains a challenging task. (Gunning *et al.*, 2021; Cheng *et al.*, 2023). Therefore, we only applied ClinGen- and ACMG-recommended tools(Waring *et al.*, 2021; Pejaver *et al.*, 2022; Wilcox *et al.*, 2022). This approach enabled us to use more reliable tools in our study. Second, due to the lack of research on MEFV gene

classification, we had to base our sample size calculations on the study by Accetturo et al. (Accetturo *et al.*, 2020). However, we also confirmed this 'optimum number of features' by looking at the literature and their classification accuracy (Vu and Braga-Neto, 2009; Accetturo *et al.*, 2020).

The main drawback of our study is the absence of validation via clinical or functional studies. While integrating our model with the ClinVar dataset could provide an avenue for external validation, ClinVar currently reports only 33 missense variants (Accessed: 12/5/2024). As we have already integrated all these variants into our training dataset, we could not utilize them as an external validation dataset. However, the explicit methodology of the modified hard voting classifier facilitates its straightforward application to diverse datasets. Further studies are necessary to fully understand the efficacy of the modified hard voting classifier.

5. Conclusion

Brief Summary of Findings and Evaluation of the study

This study holds significance for both machine learning applications and routine clinical practice. In this work, an algorithm was developed to enhance the performance of hard voting classifiers, demonstrating optimal results even with small sample sizes. However, the primary limitation of the study is that it has not been validated on an external dataset.

Consequently, this approach addresses the three previously identified gaps in in silico tools, reduces existing prediction errors of other in silico tools by offering gene-specific optimization, and, most importantly, provides an alternative method for bioinformaticians working on in silico tool optimization while also serving as a helpful tool for clinicians. Given that 60% of the clinical implications associated with MEFV gene variants are still incompletely understood, it would be advantageous to apply a modified hard voting vote classifier approach to enhance the classification accuracy of machine learning techniques. However, more testing of the improved modified hard voting approach is required on other gene variations.

Future Implications

The impact of this modified hard voting classifier on other datasets also needs to be evaluated to better understand its significance compared to the standard hard voting classifier. Additionally, from a clinical perspective, functional studies specifically designed for the *MEFV* gene are required to fully comprehend the true success of these classifications.

6. References

Accetturo, M. *et al.* (2020) 'Improvement of MEFV gene variants classification to aid treatment decision making in familial Mediterranean fever.', *Rheumatology (Oxford, England)*, 59(4), pp. 754–761. Available at: https://doi.org/10.1093/rheumatology/kez332.

Accetturo, M., Bartolomeo, N. and Stella, A. (2020) 'In-silico Analysis of NF1 Missense Variants in ClinVar: Translating Variant Predictions into Variant Interpretation and Classification.', *International journal of molecular sciences*, 21(3). Available at: https://doi.org/10.3390/ijms21030721.

Acharjee, A. *et al.* (2020) 'A random forest based biomarker discovery and power analysis framework for diagnostics research', *BMC Medical Genomics*, 13(1), p. 178. Available at: https://doi.org/10.1186/s12920-020-00826-6.

Adzhubei, I., Jordan, D.M. and Sunyaev, S.R. (2013) 'Predicting functional effect of human missense mutations using PolyPhen-2.', *Current protocols in human genetics*, Chapter 7, p. Unit7.20. Available at: https://doi.org/10.1002/0471142905.hg0720s76.

Alay, M.T. (2024) 'An Ensemble Model Based on Combining BayesDel and Revel Scores Indicates Outstanding Performance: Importance of Outlier Detection and Comparison of Models', *Cerrahpasa Medical Journal*, 48(2), pp. 179–184.

Albaradei, S. *et al.* (2021) 'Machine learning and deep learning methods that use omics data for metastasis prediction.', *Computational and structural biotechnology journal*, 19, pp. 5008–5018. Available at: https://doi.org/10.1016/j.csbj.2021.09.001.

Awe, O.O. *et al.* (2024) 'Weighted hard and soft voting ensemble machine learning classifiers: Application to anaemia diagnosis', in *Sustainable Statistical and Data Science Methods and Practices: Reports from LISA 2020 Global Network, Ghana, 2022*. Springer, pp. 351–374.

Burdon, K.P. *et al.* (2022) 'Specifications of the ACMG/AMP variant curation guidelines for myocilin: Recommendations from the clingen glaucoma expert panel.', *Human mutation*, 43(12), pp. 2170–2186. Available at: https://doi.org/10.1002/humu.24482.

Cheng, J. *et al.* (2023) 'Accurate proteome-wide missense variant effect prediction with AlphaMissense.', *Science (New York, N.Y.)*, 381(6664), p. eadg7492. Available at: https://doi.org/10.1126/science.adg7492.

Dalmaijer, E.S., Nord, C.L. and Astle, D.E. (2022) 'Statistical power for cluster analysis', *BMC Bioinformatics*, 23(1), pp. 1–28. Available at: https://doi.org/10.1186/s12859-022-04675-1.

Dundar, M. *et al.* (2022) 'Clinical and molecular evaluation of MEFV gene variants in the Turkish population: a study by the National Genetics Consortium.', *Functional & integrative genomics*, 22(3), pp. 291–315. Available at: https://doi.org/10.1007/s10142-021-00819-3.

El-Sofany, H., Bouallegue, B. and El-Latif, Y.M.A. (2024) 'A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method.',

Scientific reports, 14(1), p. 23277. Available at: https://doi.org/10.1038/s41598-024-74656-2.

Fortuno, C. *et al.* (2018) 'Improved, ACMG-compliant, in silico prediction of pathogenicity for missense substitutions encoded by TP53 variants.', *Human mutation*, 39(8), pp. 1061–1069. Available at: https://doi.org/10.1002/humu.23553.

Van Gijn, M.E. *et al.* (2018) 'New workflow for classification of genetic variants' pathogenicity applied to hereditary recurrent fevers by the International Study Group for Systemic Autoinflammatory Diseases (INSAID).', *Journal of medical genetics*, 55(8), pp. 530–537. Available at: https://doi.org/10.1136/jmedgenet-2017-105216.

Grandemange, S. *et al.* (2011) 'The regulation of MEFV expression and its role in health and familial Mediterranean fever', *Genes & Immunity*, 12(7), pp. 497–503. Available at: https://doi.org/10.1038/gene.2011.53.

Gunning, A.C. *et al.* (2021) 'Assessing performance of pathogenicity predictors using clinically relevant variant datasets', *Journal of Medical Genetics*, 58(8), pp. 547–555. Available at: https://doi.org/10.1136/jmedgenet-2020-107003.

Harrison, S.M., Biesecker, L.G. and Rehm, H.L. (2019) 'Overview of Specifications to the ACMG/AMP Variant Interpretation Guidelines.', *Current protocols in human genetics*, 103(1), p. e93. Available at: https://doi.org/10.1002/cphg.93.

Hu, Y.-H. *et al.* (2024) 'A novel MissForest-based missing values imputation approach with recursive feature elimination in medical applications', *BMC Medical Research Methodology*, 24(1), p. 269. Available at: https://doi.org/10.1186/s12874-024-02392-2.

Ioannidis, N.M. *et al.* (2016) 'REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants.', *American journal of human genetics*, 99(4), pp. 877–885. Available at: https://doi.org/10.1016/j.ajhg.2016.08.016.

Khalid, Z. and Sezerman, O.U. (2018) 'Computational drug repurposing to predict approved and novel drug-disease associations', *Journal of Molecular Graphics and Modelling*, 85, pp. 91–96. Available at: https://doi.org/https://doi.org/10.1016/j.jmgm.2018.08.005.

Kırnaz, B., Gezgin, Y. and Berdeli, A. (2022) 'MEFV gene allele frequency and genotype distribution in 3230 patients' analyses by next generation sequencing methods.', *Gene*, 827, p. 146447. Available at: https://doi.org/10.1016/j.gene.2022.146447.

Knecht, C. *et al.* (2017) 'IMHOTEP-a composite score integrating popular tools for predicting the functional consequences of non-synonymous sequence variants.', *Nucleic acids research*, 45(3), p. e13. Available at: https://doi.org/10.1093/nar/gkw886.

Lai, A. *et al.* (2022) 'The ClinGen Brain Malformation Variant Curation Expert Panel: Rules for somatic variants in AKT3, MTOR, PIK3CA, and PIK3R2.', *Genetics in medicine : official journal of the American College of Medical Genetics*, 24(11), pp. 2240–2248. Available at: https://doi.org/10.1016/j.gim.2022.07.020.

Larracy, R., Phinyomark, A. and Scheme, E. (2021) 'Machine learning model validation for early stage studies with small sample sizes', in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, pp. 2314–2319.

Liu, C. *et al.* (2013) 'Applications of machine learning in genomics and systems biology.', *Computational and mathematical methods in medicine*, p. 587492. Available at: https://doi.org/10.1155/2013/587492.

Liu, X. *et al.* (2016) 'dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs.', *Human mutation*, 37(3), pp. 235–241. Available at: https://doi.org/10.1002/humu.22932.

Luan, J. *et al.* (2020) 'The predictive performances of random forest models with limited sample size and different species traits', *Fisheries Research*, 227, p. 105534. Available at: https://doi.org/https://doi.org/10.1016/j.fishres.2020.105534.

Megantara, A.A. and Ahmad, T. (2021) 'A hybrid machine learning method for increasing the performance of network intrusion detection systems', *Journal of Big Data*, 8(1). Available at: https://doi.org/10.1186/s40537-021-00531-w.

Mighton, C. *et al.* (2022) 'Data sharing to improve concordance in variant interpretation across laboratories: results from the Canadian Open Genetics Repository', *Journal of Medical Genetics*, 59(6), pp. 571 LP – 578. Available at: https://doi.org/10.1136/jmedgenet-2021-107738.

Ng, P.C. and Henikoff, S. (2003) 'SIFT: Predicting amino acid changes that affect protein function.', *Nucleic acids research*, 31(13), pp. 3812–3814. Available at: https://doi.org/10.1093/nar/gkg509.

Nykamp, K. *et al.* (2017) 'Sherloc: a comprehensive refinement of the ACMG-AMP variant classification criteria.', *Genetics in medicine : official journal of the American College of Medical Genetics*, 19(10), pp. 1105–1117. Available at: https://doi.org/10.1038/gim.2017.37.

Ogundimu, E.O., Altman, D.G. and Collins, G.S. (2016) 'Adequate sample size for developing prediction models is not simply related to events per variable.', *Journal of clinical epidemiology*, 76, pp. 175–182. Available at: https://doi.org/10.1016/j.jclinepi.2016.02.031.

Palanivinayagam, A. and Damaševičius, R. (2023) 'Effective Handling of Missing Values in Datasets for Classification Using Machine Learning Methods', *Information*, 14(2), p. 92.

Papin, S. *et al.* (2007) 'The SPRY domain of Pyrin, mutated in familial Mediterranean fever patients, interacts with inflammasome components and inhibits proIL-1beta processing.', *Cell death and differentiation*, 14(8), pp. 1457–1466. Available at: https://doi.org/10.1038/sj.cdd.4402142.

Pejaver, V. *et al.* (2022) 'Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria.', *American journal of human genetics*, 109(12), pp. 2163–2177. Available at: https://doi.org/10.1016/j.ajhg.2022.10.013.

Pyeritz, R.E. and for the Professional Practice and

Guidelines Committee, A. (2012) 'Evaluation of the adolescent or adult with some features of Marfan syndrome', *Genetics in Medicine*, 14(1), pp. 171–177. Available at: https://doi.org/10.1038/gim.2011.48.

Rajput, D., Wang, W.-J. and Chen, C.-C. (2023) 'Evaluation of a decided sample size in machine learning applications', *BMC Bioinformatics*, 24(1), p. 48. Available at: https://doi.org/10.1186/s12859-023-05156-9.

Richards, S. *et al.* (2015) 'Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology.', *Genetics in medicine : official journal of the American College of Medical Genetics*, 17(5), pp. 405–424. Available at: https://doi.org/10.1038/gim.2015.30.

Riley, R.D. *et al.* (2019) 'Minimum sample size for developing a multivariable prediction model: PART II-binary and time-to-event outcomes', *Statistics in medicine*, 38(7), pp. 1276–1296.

Sallah, S.R. *et al.* (2022) 'Improving the clinical interpretation of missense variants in X linked genes using structural analysis.', *Journal of medical genetics*, 59(4), pp. 385–392. Available at: https://doi.org/10.1136/jmedgenet-2020-107404.

Savige, J. *et al.* (2021) 'Consensus statement on standards and guidelines for the molecular diagnostics of Alport syndrome: refining the ACMG criteria.', *European journal of human genetics : EJHG*, 29(8), pp. 1186–1197. Available at: https://doi.org/10.1038/s41431-021-00858-1.

Song, X. *et al.* (2021) 'Comparison of machine learning and logistic regression models in predicting acute kidney injury: A systematic review and meta-analysis.', *International journal of medical informatics*, 151, p. 104484. Available at: https://doi.org/10.1016/j.ijmedinf.2021.104484.

Stewart, D.R. *et al.* (2018) 'Care of adults with neurofibromatosis type 1: a clinical practice resource of the American College of Medical Genetics and Genomics (ACMG)', *Genetics in Medicine*, 20(7), pp. 671–682. Available at: https://doi.org/10.1038/gim.2018.28.

Tian, Y. *et al.* (2019) 'REVEL and BayesDel outperform other in silico meta-predictors for clinical variant classification', *Scientific Reports*, 9(1), p. 12752. Available at: https://doi.org/10.1038/s41598-019-49224-8.

Vabalas, A. *et al.* (2019) 'Machine learning algorithm validation with a limited sample size.', *PloS one*, 14(11), p. e0224365. Available at: https://doi.org/10.1371/journal.pone.0224365.

Vu, T.T. and Braga-Neto, U.M. (2009) 'Is bagging effective in the classification of small-sample genomic and proteomic data?', *EURASIP journal on bioinformatics & systems biology*, 2009(1), p. 158368. Available at: https://doi.org/10.1155/2009/158368.

Waring, A. *et al.* (2021) 'Data-driven modelling of mutational hotspots and in silico predictors in hypertrophic cardiomyopathy.', *Journal of medical genetics*, 58(8), pp. 556–564. Available at: https://doi.org/10.1136/jmedgenet-2020-106922.

Wilcox, E.H. *et al.* (2022) 'Evaluating the impact of in silico predictors on clinical variant classification.', *Genetics in medicine : official journal of the American College of Medical Genetics*, 24(4), pp. 924–930. Available at: https://doi.org/10.1016/j.gim.2021.11.018.