Cerrahpaşa Med J 2024; 48(2): 179-184 ORIGINAL ARTICLE

An Ensemble Model Based on Combining BayesDel and Revel Scores Indicates Outstanding Performance: Importance of Outlier Detection and Comparison of Models

Mustafa Tarık Alay

Department of Medical Genetics, Etlik City Hospital, Ankara, Türkiye

Cite this article as: Alay MT. An ensemble model based on combining BayesDel and Revel scores indicates outstanding performance: Importance of outlier detection and comparison of models. *Cerrahpaşa Med J.* 2024;48(2):179-184.

Abstract

Objective: Our objective is to create an effective ensemble tool that can accurately predict *MEFV* gene variants and determine the threshold value for pathogenicity based on the optimal distribution.

Methods: First, we extracted a dataset from the Infevers database [https://infevers.umai-montpellier.fr/web/search.php?n=1]. Second, we merged the variant classification into 2 categories: likely benign and likely pathogenic. Third, we implemented our high-sensitivity model to obtain disease-causing variants. In the 4 steps, we implemented curve estimation analysis to determine which curve was fitting our variant distribution. We implemented the receiver operating curve after the curve estimation analysis to find suitable in silico tool models for logistic regression. Repeated outlier detection analysis was performed in the fifth step until no outliers were detected. Ensemble tree-based machine-learning models were used to test a statistical model in the final step.

Results: When outliers were taken out, the Revel and BayesDel algorithms both had much higher ROCAUC scores (0.982 [0.967-0.998], $P < .001$ for the combined model; 0.982 [0.967-0.998], $P < .001$ for Revel; and 0.933 [0.889-0.977], $P < .001$ for BayesDel). AdaBoost was the most accurate machine learning model, with 0.982 ROACUAC scores.

Conclusion: Our study revealed that the implementation of outlier and anomaly detection techniques can enhance the accuracy of statistical models and yield more precise outcomes in machine learning datasets.

Keywords: Logistic regression, machine learning, MEFV, outlier, sensitivity

Introduction

Familial Mediterranean fever (FMF) is an autosomal recessively inherited disorder that primarily affects serous organs and is characterized by high fever, abdominal pain, chest pain, and arthritis.[1] *MEFV* gene mutations are responsible for the clinical manifestations of FMF. The *MEFV* gene encodes for the pyrin protein.[2] According to the International Study Group for Systemic Autoinflammatory Diseases (INSAID) consensus criteria, more than half of the *MEFV* gene variants are not classified as benign or pathogenic.[3] Therefore, it is an urgent necessity to classify unknown *MEFV* gene variants. There are 2 traditional ways that have been available during the variant classification process: (1) functional studies, and (2) clinical studies, which include a larger number of patients. Although these methods provided accurate solutions, they required a significant amount of time and money. Therefore, artificial intelligence (AI) methods, such as combinatory evaluation of in silico tools, can provide fast, accurate, and cheaper solutions.

In silico tool prediction is evaluated in supporting roles according to American College of Medical Genetics (ACMG) criteria.[4] During the variant classification process, many variant prediction tools are used. With the existing tools, the threshold classification accuracy of many tools is very low. In 2022, the Clingen Group recommended a new threshold for certain protein prediction tools and meta-predictors. However, they stressed that their recommended thresholds could vary from gene to gene. Therefore, they suggested that specific groups are needed to determine gene-specific thresholds.[5] In certain conditions, a laboratory with a specific gene or a limited number of genes as its focus may autonomously calibrate one of these tools using the methodology outlined in Clingen's recommendation.

The calibration process may result in unique numerical thresholds for different levels of evidence specifically tailored to the gene(s) under investigation. While most of the predictors classified benign *MEFV* gene variants with higher accuracy, they could not classify pathogenic variants even with heads-or-tails probability.[5,6] Furthermore, most of the time, the predictors underestimate the variations of unknown significance and their effects. Therefore, the prediction calculations might yield accuracy scores that are overestimated. However, including variation of unknown significance (VOUS) variants in prediction calculations might provide 2 advantages: (1) increased sample size and (2) minimized the effects of selection bias. Because we are facing a serious sensitivity problem, optimal thresholds are needed for solving this problem. The combination of the optimal number of tools might provide essential benefits for variant classification.

In this research, *Z*-score-based outlier analysis is utilized. The reason for choosing *Z*-scores is their status as one of the most commonly used methods for outlier analysis. However, the

Received: January 3, 2024 **Revision Requested:** February 5, 2024
Last Revision Received: February 10, 2024 **Accepted:** February 10, 2024
Publication Date: July 10, 2024
Corresponding author: Mustafa Tarık Alay, Department of Medical Genetics, Etlik City Hospital, Ankara, Türkiye
e-mail: mtarikalay@gmail.com
DOI: 10.5152/cjm.2024.23124

Content of this journal is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

difference here lies in the repeated application of this analysis until no outliers remain. This approach aims to create a dataset with a distribution as close to normal as possible, minimizing data loss. We will establish new pathogenicity threshold values based on this dataset. The study aims to develop an Ensemble tool that achieves the most accurate prediction possible and determines the pathogenicity threshold value provided by the most optimum distribution.

Methods

Data Extraction and Preparation for Analysis

First, we extracted a dataset from the Infevers database [https ://infevers.umai-montpellier.fr/web/search.php?n=1. We obtained common variant prediction tools from the dbNFSP 4.0 and Franklin Genoox databases. All prediction tools were selected according to Clingen's recommendations.[4,5,7] Aggregated prediction scores from Franklin Genoox were used as a control group. Franklin Genoox is a well-respected variant interpretation tool such as Varsome and Intervar.[8] Therefore, we used aggregated predictions as a control group for our receiver operating curve analysis. Overall, 266 out of 389 (68.38%) variants were obtained, and 11 out of 43 (25.58%) scores were available for analysis, which are Bayes Del, FATHMM, GeneCanyon, GERP, fitCons, MetaLR, MutAssesor, Polyphen-2, Revel, SIFT, and Varity. These tools were selected based on Clingen Research Group in silico tool selection recommendations for missense variants.[5] Second, we included only single nucleotide polymorphism variants, whether located in the coding region or not. Hence, we excluded frameshift deletions, inframe deletions, frameshift insertions, inframe insertions, and duplications.The study relies on publicly available and open-source data. There is no requirement for approval from an ethics committee or informed consent.

Variant Categorization

In the Infevers database, 7 categories are available: not classified, unsolved, variation of unknown significance, benign, likely benign, pathogenic, and likely pathogenic. According to ACMG criteria, there are 5 classification categories: variation of unknown significance, benign, likely benign, pathogenic, and likely pathogenic. However, in routine clinical practice, likely pathogenic and pathogenic variants are most commonly used to describe disease-causing variants.[4,5,9] Furthermore, likely benign and benign variants are also utilized to describe benign variants. We use unresolved, VOUS, and not categorized to describe indeterminate variants. We merged the likely pathogenic and pathogenic variants as disease-causing in the high-sensitivity group and coded them as 1; we named other variants as "others" and coded them as 0. Similarly, we merged the likely benign and benign variants as benign in the high specificity group and assigned them a code of 1; all other variants were named and coded as 0. A similar merged method was used in previous studies.[6,10,11]

High Sensitivity Model

The concept of sensitivity is commonly defined as the proportion of true positive results to the overall number of individuals who truly possess the specific condition under examination. In contrast, specificity can be defined as the proportion of true negatives relative to the overall population of individuals who do not possess the specific condition under consideration for testing. The evaluation of diagnostic procedures necessitates the implementation of these measures, as they play a pivotal role in determining their accuracy and reliability.[12] Therefore, disease-causing and benign variants are target variables in high-sensitivity and high-specificity models, respectively.

Curve Estimation Analyses

The primary aim of curve fitting is to provide a theoretical representation of empirical data using a model, typically in the form of a function or equation, and to determine the corresponding parameters for this model. Mechanistic models hold paramount significance in our context. Curve estimation analyses are commonly employed to determine whether a linear or logistic model adequately fits a given dataset. These analyses aim to establish the presence of a statistically significant curve distribution.[13] However, our analysis encompassed not only linear models but also quadratic, logistic, and other types of models.

Receiver Operating Curve Analysis

Based on the receiver operating characteristic (ROC) analysis, scores ranging from 0.7 to 0.8 are generally considered acceptable, while scores falling between 0.8 and 0.9 are deemed more favorable. Exceeding 0.9, scores are regarded as indicative of outstanding performance. We established a threshold of 0.7. For the purpose of prediction, scores exceeding the threshold of 0.7 are utilized. Logistic regression (LR) analysis incorporates scores that exceed the threshold of 0.7.

Logistic Regression Model for Prediction

The high-sensitivity model of the LR model is predicted to find disease-causing variants, while the high-specificity model is predicted to find benign variants. Before implementing LR analysis, we checked the necessary assumptions for implementing LR analysis. These include multicollinearity problems, homogeneity of variance, and outlier detection.

Outlier Detection

We repeated the outliers detection process until there were no outliers remaining within the range between –2 and +2 SD, as determined by the *Z*-score distribution.

Machine Learning Classification Process

We outperformed LR, Gradient Boosting, Random Forest, and Ada Boosting algorithms. The reason that we chose LR is to test our prediction algorithm's success on a machine-learning dataset. Tree-based algorithms are widely accepted as the most accurate classifiers and are not significantly influenced by outlier detection. Therefore, we selected 3 tree-based algorithms that were not affected by outliers.

Results

Selection of In Silico Tools and Curve Fit Analysis Results

The exclusion of Polyphen-2 from the analysis was based on the presence of more than 80 missing data points. The analysis using GeneCanyon, FATHMM, MutAssesor, SIFT, and Fitcons in silico models revealed that the curves in question did not fit with any known curve model. Figure 1 illustrates the findings.

Prediction Results on Outlier Included Analysis Results and Finding Most Accurately Classified Tools

We performed ROC analysis on the remaining Revel, MetaLR, BayesDel, Varity, and GERP scores. The analysis findings conclude that all 5 scores exhibited a statistically significant ability to identify the variants deemed causative of the disease under consideration. However, the AUC analysis revealed that 3 scores surpassed

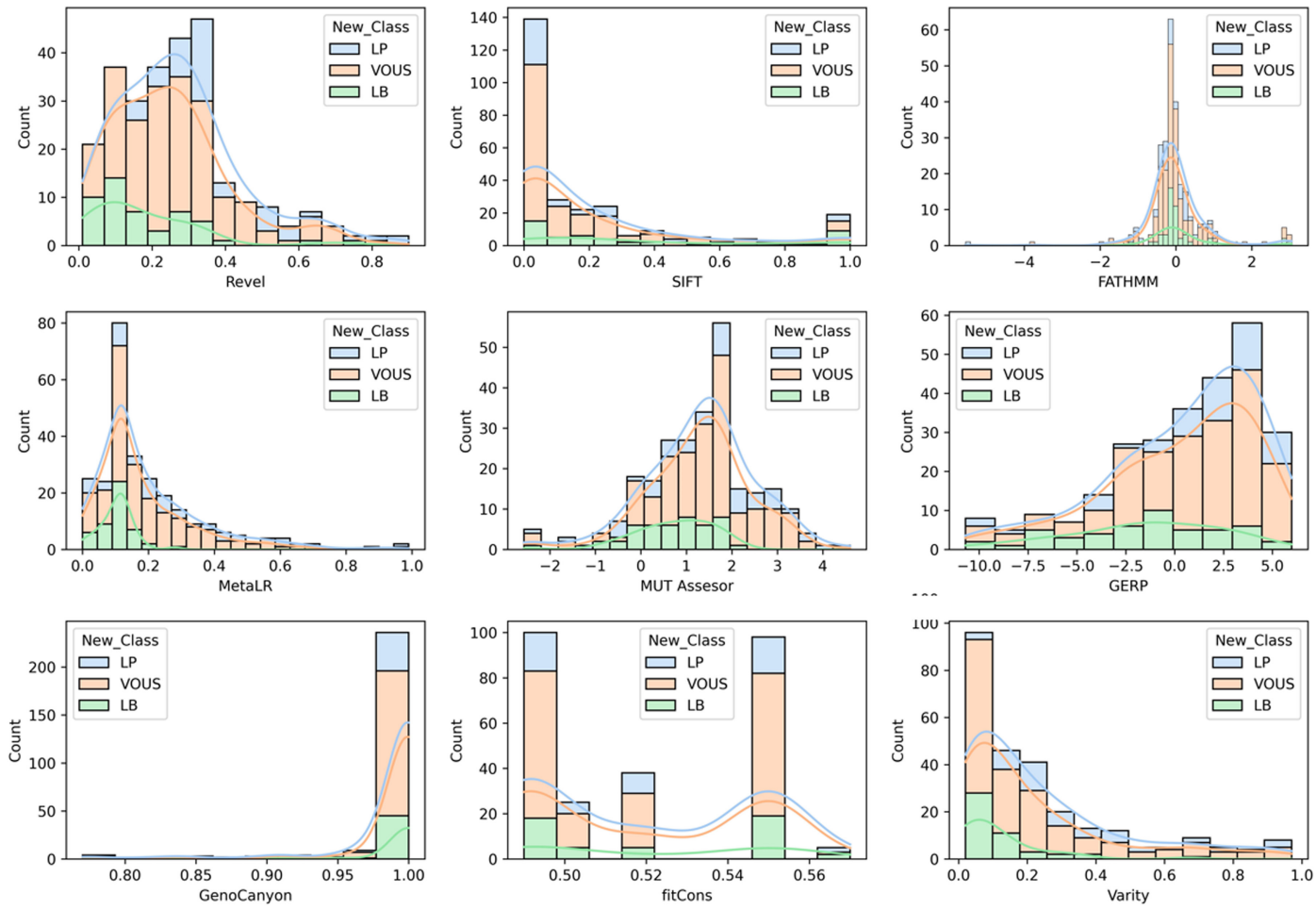


Figure 1. Unnormalized score comparison of in silico tools according to new classification system. While Revel, SIFT, MetaLR, and Varity scores have been right-skewed distributed, GERP and GenoCanyon algorithms were left-skewed.

the threshold of 0.7. We evaluated the scores Revel, BayesDel, and Varity based on their performance in detecting the area under the curve. The respective scores were as follows: 0.747 [0.680-0.814], $P < .001$; 0.743 [0.674-0.811], $P < .001$; and 0.729 [0.661-0.796], $P < .001$. The area under the curve for both the MetaLR and GERP scores was less than or equal to 0.7 (MetaLR: 0.618 [0.513-0.722], $P < .001$; GERP: 0.610 [0.521-0.699], $P < .001$). Homoscedasticity tests were conducted on the 3 scores (Revel, BayesDel, and Varity) that exceeded the set limit, revealing no multicollinearity problems (Figure 2). Subsequently, we performed LR analysis and found that the variance score was not statistically significant. There are only 2 scores left, and we combined these 2 scores as predicted probabilities and compared them with existing algorithms (Revel and BayesDel).

Implementing Repeated Analysis of *Z* Score Outlier Prediction Results and Determining New Thresholds

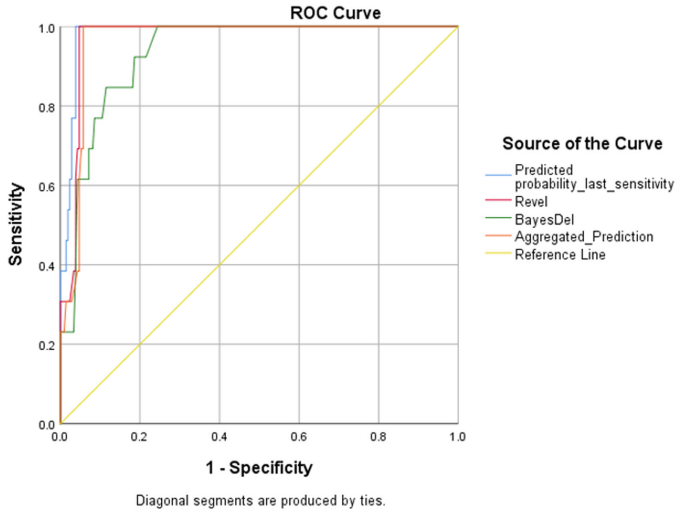
Existing algorithms include many extreme values and outliers. Therefore, we first implemented LR analysis and detected outliers according to standardized residuals. We decided to use a threshold of 2 for absolute values based on *Z* scores. After deciding on the scores as a threshold, we accepted values higher than these as outliers and excluded these variant scores from the analysis. This process repeated itself after all the outliers vanished. In the end, overall, 222 variables remained. In the final stage, we improved the combined effects of the Revel and BayesDel algorithms with outstanding performance (0.982 [0.967-0.998], $P < .001$) (Figure 3). Furthermore, each of the Revel and Bayes Del algorithms' classification success was highly increased, respectively (0.982 [0.967-0.998], $P < .001$) for Revel and 0.933 [0.889-0.977], $P < .001$) for Bayes Del. Our algorithm achieved more accurate classification than the Franklin Genoox aggregated prediction algorithms, regardless of whether outliers were removed or not. After removing outliers in our predicted probability scores higher than 0.2, scores were much more probably not classified as benign (Figure 4). Therefore, we predict that VOUS variants higher than the 0.2 threshold are most likely to be evaluated as pathogenic.

Testing Prediction Accuracy on Machine Learning Models

In the next step, we tested our outlier-detected MEFV variants on Ensemble machine-learning models. According to this model, our algorithm showed outstanding performance for each of the 4 scores, which were LR, Gradient Boosting, Random Forest, and Ada Boosting (ROCAUC >90% for every 4 algorithms). The most accurate classifier was the AdaBoost algorithm ($ROCAUC_{AdaBoost:}$ 0.9818). Although LR prediction was slightly lower than the outlier detection statistical model and other ML models, it still showed outstanding performance with 0.92 accuracy (Figure 5).

Discussion

Our model presents the best Ensemble tool to date in MEFV gene classification. Furthermore, our research calculates the optimal



Area Under the Curve

Test Result Variable(s)	Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
				Lower Bound	Upper Bound
Predicted probability_sensitivity	.982	.008	.000	.967	.998
Revel	.971	.011	.000	.950	.992
BayesDel	.933	.022	.000	.889	.977
Aggregated_Prediction	.964	.012	.000	.939	.988

Figure 3. Final step of evaluation. After removing outliers, new sensitivity scores. After repeated exclusion of outlier values, overall 44 values were excluded from the analysis, and total 222 target variations remained. Before removing outliers, the most accurate classifiers classified good (0.7-0.8 AUC values) or satisfactory (0.6-0.7 AUC values) level; however, after removing outliers the ROCAUC scores of all 4 algorithms (Combined effects of the Revel and BayesDel, Revel alone, BayesDel alone, and Franklin Genoox aggregated prediction) indicates outstanding performance (higher than 0.9 AUC values) Removing outliers not only increased significantly our predicted probability algorithm ROCAUC performance but also contributed significantly to each individual prediction tool.

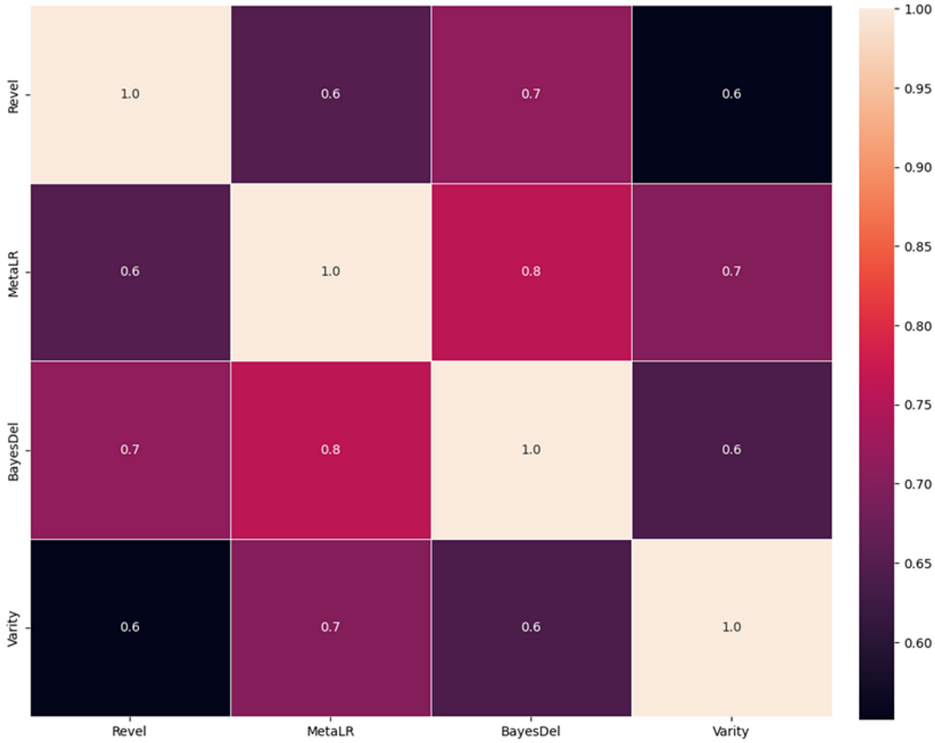


Figure 2. Correlation analysis of the 4 most accurately classified tools. Higher correlation values than 0.7 accepted as a high correlation and therefore excluded from the model. The most significant correlation was detected between MetaLR and BayesDel scores (*r*:0.8). As there were strong correlation detected between MetaLR and BayesDel, MetaLR scores were excluded from the analysis.

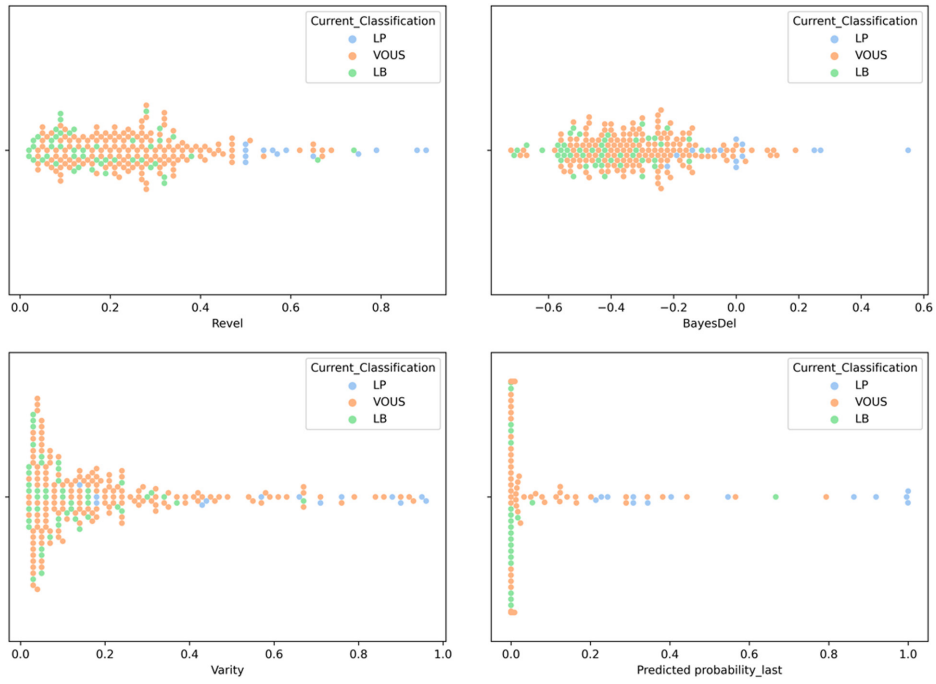


Figure 4. After the removal of outliers, the distribution of each variant was subsequently classified into a new categorization. Our predictive probability algorithms demonstrate practically optimal outcomes. Based on the outcomes of our algorithm, any score exceeding 0.2, with the exception of one value, was categorized as LP. Alternatively, when using a threshold of 0.5, the classification of variants was even worse than random chance. Conversely, scores below this threshold were classified as benign. Nevertheless, despite extensive outlier detection, the remaining 3 algorithms failed to differentiate between LP and LB variants on an individual basis.

threshold for every accurate classifier prediction tool. Our algorithm outperformed with 98.6% ROCAUC scores in classification. Moreover, the algorithm presents the most accurate classification among the current tools. The success of our Ensemble algorithm is based on including Revel and BayesDel combinations. Therefore, we both evaluated and removed the optimal number of outliers in both Revel and BayesDel.

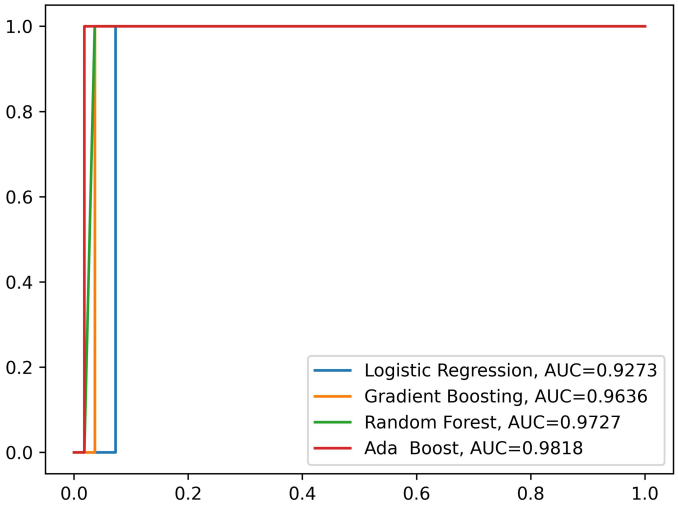


Figure 5. Results of machine learning. All 4 algorithms exhibit outstanding classification performance. The ROCAUC results of the Ensemble algorithm were found to be higher than those of logistic regression. Despite the logistic regression algorithm's prediction scores being lower than anticipated (ROCUAC:0.92), all algorithms exhibited exceptional performance on the test dataset. The classifier that achieved the highest level of accuracy was AdaBoost, with a classification success rate of 98%.

According to our study results, only 2 variant pathogenicity predictor tools are better than Franklin Genoox's aggregated prediction, which are BayesDel and Revel, compatible with the existing literature.[5,14] Surprisingly, the BayesDel algorithm presents a more accurate classification ratio than the Revel algorithm. However, after removing outliers, Revel outperformed in classification. Currently, numerous tools predict missense variants. A few months ago, alpha missense predicted all missense variant combinations based on alpha fold, which predicts alpha fold 3D visualization prediction.[15,16] Although the pathogenicity of more than 90% of missense variants is unknown in the ClinVar database,[15] alpha missense can only be classified in more than 60% of all variants.[16] However, despite the significant improvements in technologies, we are still far beyond the classification of many unknown variants. Furthermore, with the widespread utilization of next-generation sequencing technology, we are facing novel, uncategorized variants.[17]

It is quite certain that novel technologies bring new, unknown issues. However, solving these issues can be achieved by combining novel technologies, novel statistical approaches, and optimizing existing algorithms. In the literature, few studies aimed to optimize existing scores for MEFV gene variants. One of the optimization studies on MEFV gene variants was implemented by determining a new threshold by Acetturo et al; however, they tested their results on relatively low training dataset accuracy,[18] which is 75% accuracy.[11] In this article, they included an analysis of outliers and testing for training data accuracy. They used linear discriminant analysis (LDA). However, because the dataset lacks a Gaussian distribution, LDA is inappropriate for their analysis. Therefore, the assumption of equal covariance matrices across classes might not hold true.[19] In contrast to this study, we used LR

analysis[20] in our analysis, and we used curve fitting analysis in the previous step to test the distributions of our datasets.[21,22] However, the new classifier algorithms implemented by Alay et al, dubbed "modified hard margin classifiers," are able to classify all likely pathogenic and likely benign variants correctly. To find out how accurate the data is in reality, the algorithm must test it on additional datasets.[6]

The strength of the study was based on optimum, most accurate classifiers and outlier-detected methods. The limitation of our study is the sample size based on the reported number of *MEFV* gene variants. However, the sample size is sufficient to implement outlier detection. Furthermore, Ensemble machine learning methods' outstanding classification success supported classical statistical method concepts. The algorithm can be based on easily repeatable and commonly accepted statistical concepts, and it can be easily adapted to other gene classification approaches. Therefore, using the same algorithm, we can predict other genes. In the next step, the outlier-detected Ensemble method can be implemented for other gene groups. Although algorithmic classification improvement is quite assured, the classification accuracy of in silico tools can vary from gene to gene. Therefore, it is imperative to implement these new methods on other genes for 2 reasons: (1) determining the best in silico predictors of every gene and detecting gene-specific thresholds for each gene and (2) testing the generalizability of the algorithm.

Conclusion

As a result of the study, we detected that outlier and anomaly detection can improve not only statistical model accuracy but also provide more accurate results in machine learning datasets. Therefore, we strongly recommend that determining thresholds be based on outliers removal. The recommendation is based on 2 issues: (1) outlier values can cause misinterpretation and decrease accuracy. (2) Outliers can reduce threshold determination. Our approach has the capacity to significantly improve the precision of categorizing MEFV gene variants and establish a more precise and accurate threshold by reducing the influence of outliers on classification bias. We can also utilize this approach for unknown variants in alternative genes.

Data Availability Statement: The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Ethics Committee Approval: All data were extracted from the nfevers database. The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Informed Consent: The study is based on open-source data. No informed consent is required.

Peer-review: Externally peer-reviewed.

Acknowledgments: The author thanks Dr. İsmail Alay for his helping in the data collection step.

Declaration of Interests: The author has no conflict of interest to declare.

Funding: The author declared that this study has received no financial support.

References

1. Yıldız M, Haşlak F, Adrovic A, Barut K, Kasapçopur Ö. Autoinflammatory diseases in childhood. *Balk Med J*. 2020;37(5):236-246. [CrossRef]
2. Dundar M, Fahrioglu U, Yildiz SH, et al. Clinical and molecular evaluation of MEFV gene variants in the Turkish population: a study by the National Genetics Consortium. *Funct Integr Genomics*. 2022;22(3):291-315. [CrossRef]
3. Van Gijn ME, Ceccherini I, Shinar Y, et al. New workflow for classification of genetic variants' pathogenicity applied to hereditary recurrent fevers by the International Study Group for Systemic Autoinflammatory Diseases (INSAID). *J Med Genet*. 2018;55(8):530-537. [CrossRef]
4. Richards S, Aziz N, Bale S, 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(5):405-424. [CrossRef]
5. Pejaver V, Byrne AB, Feng BJ, et al. Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. *Am J Hum Genet*. 2022;109(12):2163-2177. [CrossRef]
6. Alay MT, Demir İ, Kirişçi M. *Three Steps Novel Machine Learning Method Classifies Uncertain MEFV Gene Variants*. Published Online 2023.
7. Burdon KP, Graham P, Hadler J, et al. Specifications of the ACMG/AMP variant curation guidelines for myocilin: recommendations from the clingen glaucoma expert panel. *Hum Mutat*. 2022;43(12):2170-2186. [CrossRef]
8. Mighton C, Smith AC, Mayers J, et al. Data sharing to improve concordance in variant interpretation across laboratories: results from the Canadian Open Genetics Repository. *J Med Genet*. 2022;59(6):571-578. [CrossRef]
9. Nykamp K, Anderson M, Powers M, et al. Sherloc: a comprehensive refinement of the ACMG-AMP variant classification criteria. *Genet Med*. 2017;19(10):1105-1117. [CrossRef]
10. Accetturo M, Bartolomeo N, Stella A. In-silico analysis of NF1 missense variants in ClinVar: translating variant predictions into variant interpretation and classification. *Int J Mol Sci*. 2020;21(3). [CrossRef]
11. Accetturo M, D'Uggento AM, Portincasa P, Stella A. Improvement of MEFV gene variants classification to aid treatment decision making in familial Mediterranean fever. *Rheumatol (Oxf Engl)*. 2020;59(4):754-761. [CrossRef]
12. Monaghan TF, Rahman SN, Agudelo CW, et al. Foundational statistical principles in medical research: sensitivity, specificity, positive predictive value, and negative predictive value. *Medicina (Kaunas)*. 2021;57(5). [CrossRef]
13. Gonçalves L, Subtil A, Oliveira MR, de Zea Bermudez P. ROC curve estimation: an overview. *REVSTAT Stat J*. 2014;12(1):1-20.
14. Tian Y, Pesaran T, Chamberlin A, et al. REVEL and BayesDel outperform other in silico meta-predictors for clinical variant classification. *Sci Rep*. 2019;9(1):12752. [CrossRef]
15. Schmidt A, Röner S, Mai K, Klinkhammer H, Kircher M, Ludwig KU. Predicting the pathogenicity of missense variants using features derived from AlphaFold2. *Bioinformatics*. 2023;39(5). [CrossRef]
16. Cheng J, Novati G, Pan J, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*. 2023;381(6664):eadg7492. [CrossRef]
17. Hu T, Chitnis N, Monos D, Dinh A. Next-generation sequencing technologies: an overview. *Hum Immunol*. 2021;82(11):801-811. [CrossRef]
18. Polo TCF, Miot HA. Use of ROC curves in clinical and experimental studies. *J Vasc Bras*. 2020;19:e20200186. [CrossRef]
19. Hastie T, Tibshirani R. Discriminant analysis by Gaussian mixtures. *J R Stat Soc B*. 1996;58(1):155-176. [CrossRef]
20. LaValley MP. Logistic regression. *Circulation*. 2008;117(18):2395-2399. [CrossRef]
21. Levie R de. Curve fitting with least squares. *Crit Rev Anal Chem*. 2000;30(1):59-74. [CrossRef]
22. Uhler HS. Method of least squares and curve fitting. *J Opt Soc Am*. 1923;7(11):1043-1066. [CrossRef]