Original Article

MEDICINE PALLIATIVE CARE

DOI: 10.47582/jompac.1259507

J Med Palliat Care 2023; 4(2): 117-125

How to explain a machine learning model: HbA1c classification example

Deniz İlhan Topçu

Department of Medical Biochemistry, İzmir Tepecik Training and Research Hospital, University of Health Sciences, İzmir, Turkey

Cite this article as: Topcu Dİ. How to explain a machine learning model: HbA1c classification example. J Med Palliat Care 2023; 4(2): 117-125.

ABSTRACT

Aim: Machine learning tools have various applications in healthcare. However, the implementation of developed models is still limited because of various challenges. One of the most important problems is the lack of explainability of machine learning models. Explainability refers to the capacity to reveal the reasoning and logic behind the decisions made by AI systems, making it straightforward for human users to understand the process and how the system arrived at a specific outcome. The study aimed to compare the performance of different model-agnostic explanation methods using two different ML models created for HbA1c classification.

Material and Method: The H_2O AutoML engine was used for the development of two ML models (Gradient boosting machine (GBM) and default random forests (DRF)) using 3,036 records from NHANES open data set. Both global and local model-agnostic explanation methods, including performance metrics, feature important analysis and Partial dependence, Breakdown and Shapley additive explanation plots were utilized for the developed models.

Results: While both GBM and DRF models have similar performance metrics, such as mean per class error and area under the receiver operating characteristic curve, they had slightly different variable importance. Local explainability methods also showed different contributions to the features.

Conclusion: This study evaluated the significance of explainable machine learning techniques for comprehending complicated models and their role in incorporating AI in healthcare. The results indicate that although there are limitations to current explainability methods, particularly for clinical use, both global and local explanation models offer a glimpse into evaluating the model and can be used to enhance or compare models.

Keywords: Machine learning, explainable artificial intelligence, glycated hemoglobin

INTRODUCTION

In healthcare, machine-learning (ML) models are used for various tasks, such as image and signal analysis, disease diagnosis, treatment planning, and drug discovery (1). The use of ML models to improve patient care is a novel approach, but its implementation in clinical practice is still limited (2).

Explainability can be defined as the capability of making the decision-making process of AI systems transparent and understandable for human users, it includes how the decision was reached and how the system arrived at a particular conclusion (3). It is one of the most important limitations, and it has long been a question of great interest in a wide range of fields, including medicine (4). More recently, there has been growing number of publications that focus on explainable artificial intelligence (xAI) (5).

The debate continues regarding the best strategies for the appropriate application of explainability tools. To date, there has been little agreement about what constitutes sufficient explainability and xAI suffers from insufficient application, and limited studies have investigated whether xAI contributes to ML model use in medicine(4,6). Model-agnostic explainability techniques refers to methods or techniques that can be applied to any model, regardless of its architecture or learning algorithm, and these methods can be further categorized into local and global methods. While global interpretability focuses on understanding the overall functioning and decision-making processes of a model, local interpretability focuses on understanding the reasoning behind individual predictions made by the model (5,7).

Diabetes mellitus is a chronic medical condition characterized by high blood sugar levels resulting from defects in insulin production, insulin action, or both. Monitoring HbA1c levels is critical for managing diabetes and preventing complications such as kidney damage, nerve damage, and cardiovascular disease (8). The aim of this study is to develop an ML model that utilizes

Corresponding Author: Deniz Topcu, ditopcu@gmail.com

Received: 23.02.2023 **Accepted:** 15.03.2023

routine laboratory and clinical data for hemoglobin A1c prediction and investigate the effectiveness of cutting-edge global and local explainability tools that are used for prediction explanations by comparing different models. To achieve this goal, two different ML algorithms were developed and applied to both local and global explainability model predictions.

MATERIAL AND METHOD

Data Source and HbA1c Classification

All procedures were carried out in accordance with the ethical rules and the principles of the Declaration of Helsinki. Open data sets were utilized for the study. Therefore, ethics committee approval was not obtained. The National Health and Nutrition Examination Survey data sets between 2014–2017 were used (9). Only records of adults aged 18 years and older were included in the study. Twenty-eight parameters, including clinical laboratory results and clinical information, were selected for ML model development. Regarding prediction, quantitative hemoglobin A1c (A1c) results were split into the following three classes according to the criteria recommended by the American Diabetes Association: normal (<5.7%), prediabetes (5.7%–6.4%), and diabetes (>6.4%) (10). The descriptive statistics of the included parameters are given in **Table 1**.

Machine-Learning Model Development

Pre-process, data cleaning, and generation of training and test sets: Data preparation, the creation of ML models, and statistical analyses were all completed using R statistical software version 4.2 (11). The model explanations were carried out using the H_2O and DALEX packages (12,13). The data set was composed of 3,036 records that contained all selected parameters. Therefore, there were no missing values in the data set. The data were split into training (70%, n=2,114) and test sets (30%, n=912) using stratification according to A1c, age, and sex.

Table 1. Clinical laboratory results and demographic features of the study population and feature importance for the developed models

Parameter	n (%)	Min	Max	Mean (SD)	Median (IQR)	Feature Importance1	
						ML Model	
						DRF	GBM
Age		18	80	44.7 (17.8)	41 (30)	0.36*	0.15*
ACR, Urine, mg/g		0.94	7980	39.1 (257)	6.54 (8.07)	0.07	0.01
Albumin, g/L		21	54	42.1 (3.59)	42 (5)	0.08	0.01
ALP, U/L		23	347	71.9 (24)	69 (26)	0.06	0.01
ALT, U/L		3	181	25 (16.5)	21 (14)	0.06	0.02
AST, U/L		8	289	24 (13.4)	21 (8)	0.04	0.01
BMI kg/m²		15.5	65.3	29.1 (6.9)	28 (8.5)	0.10	0.02
BUN, mg/dL		2	79	14.4 (5.63)	14 (6)	0.07	0.00
Calcium, mg/dL		7.8	10.5	9.27 (0.329)	9.3 (0.5)	0.06	0.00
Cholesterol, mg/dL		76	433	182 (39.6)	178 (52)	0.09	0.03
Creatinine, mg/dL		0.3	6.73	0.91 (0.29)	0.88 (0.29)	0.08	0.01
Glucose, mg/dL		19	434	104 (34.9)	95 (17)	0.99*	0.98*
HDL-C, mg/dL		6	151	51.6 (15.2)	49 (19)	0.08	0.01
Hemoglobin g/dL		6.3	19	14.5 (1.55)	14.6 (2)	0.09	0.02
Lymphocyte, %		7	94.5	31.2 (8.64)	30.7 (11.5)	0.07	0.01
MCV, fL		51.6	114	88.6 (6.19)	89.1 (6.6)	0.09	0.05*
Neutrophil, %		3.6	85.8	56.7 (9.37)	57 (12.5)	0.09	0.03
Platelet, 10^3 cells/µL		14	662	232 (61.7)	225 (75)	0.08	0.02
RBC, 10^6cells/µL		2.52	6.82	4.9 (0.50)	4.9 (0.66)	0.07	0.01
Bilirubin, mg/dL		0	2.8	0.57 (0.309)	0.5 (0.3)	0.07	0.02
Total Protein g/L		56	90	71.9 (4.25)	72 (6)	0.07	0.01
Triglyceride, mg/dL		10	2140	109 (88.5)	89 (76.2)	0.09	0.02
Uric acid, mg/dL		1.8	18	5.69 (1.47)	5.6 (2)	0.08	0.01
Waist Circ., cm		63.2	170	99.9 (17.2)	98.4 (22.5)	0.14*	0.02
WBC, 10^3 cells/µL		2.5	117	6.88 (2.85)	6.6 (2.4)	0.07	0.01
Sex						0.03	0.00
Male	2201 (72.5%)						
Female	835 (27.5%)						
HbA1c Class							
Normal	1898 (62.5%)						
Prediabetes	788 (26.0%)						
DM	350 (11.5%)						

ACR: Albumin creatinine ratio, Circ: Circumference, DRF: Distributed Random Forest, GBM: Gradient Boosting Machine, IQR: Interquartile range, SD: Standard deviation, 1 Scaled Importance, * Top three features

To identify the important features for developing a ML model, the study utilized the Boruta feature selection algorithm. This algorithm generated a shadow feature for each attribute by shuffling the values of the original attributes across properties. The importance of the features was then categorized into three classes: "discard" (red), "speculative" (blue), tentative (yellow) and "keep" (green) to identify the significant features (14). According to the Boruta analysis results Monocyte % parameter was excluded for ML development. Details of Boruta analysis were given in supplementary material.

Utilization of the AutoML tool: In the study, ML models were developed using the H_2O AutoML engine. H_2O is an open source, distributed ML platform that can perform all ML model development steps, including data processing, feature engineering, model building, hyperparameter optimization, and performance evaluation. The H_2O engine was utilized to develop ML models for the multinomial (multiclass) classification of A1c. Gradient boosting machine (GBM) and default random forests (DRF) models were selected as candidate algorithms. All model development-related steps were performed by the AutoML tool using the training data set. Because of the unequal distribution of A1c classes, the "balance_classes" option was used. This feature could be utilized to equalize the distribution of classes in a dataset. When activated the majority classes are either undersampled or the minority classes are oversampled. The resulting model will correct the final probabilities using a monotonic transform (12). Further, during the model development phase, hyperparameters optimization was performed using k-fold cross-validation (k=10). Finally, multiple models that were developed by the AutoML tool were evaluated using an automatically split leaderboard data set, and the winning tuned models were determined. In the study, the best models were used for the explainability method comparison.

Model explainability comparison: Model agnostic explainability methods were applied to the developed ML models. Both global and local explainability approaches were used to compare the effectiveness of the explainability methods for the DRF and GBM ML models.

Global exploration: Global methods are useful for understanding the overall patterns and behaviors of a ML model. They provide an average understanding of the model's performance. These types of methods are particularly helpful when the person building the model wants to gain a general understanding of how the model works or troubleshoot any issues with the model. In the context of the study, the following global explorations were performed to compare the internal reasoning of two ML models:

a. Performance metrics: (i) The mean per class error is the average of the errors of each class in multinominal models. It represents the misclassification of the data across the classes, and lower metrics indicate a better performance. (ii) The area under the receiver operating characteristic curve (AUC) is a metric that evaluates the model's performance for distinguishing true positives and negatives and is normally used for binominal classification problems. However, it is also possible to calculate the AUC for multinominal models using different approaches. In the study, the method that was suggested by Till was used to calculate the AUC (15). Because of the imbalanced classes, the one vs. one calculation method was used for the AUC calculation (iii) The area under the precision recall curve (AUCPR) is important and is not affected by the true negatives. Therefore, it is preferred for imbalanced data sets (15). When dealing with imbalanced data, many true negatives often make it difficult to see the impact of changes in other metrics, such as false positives. AUCPR is more responsive to changes in true positives, false positives, and false negatives than the AUC, making it a better choice for evaluating highly imbalanced data sets (12).

b. Variable importance (VIP): Variable importance is the measurement of how much each feature contributes to the ML model's predictions. This method ranks the features based on their relative importance to the model. Therefore, VIP can give a broad overview of the model's characteristics. There are several methods to measure variable importance, such as permutation importance and importance based on Shapley additive explanations (SHAP) values (16, 17). In the study, VIP values were calculated for all parameters. The significance of each variable is determined by evaluating the relative impact of the variable in the tree-building process based on its frequency of being chosen as a splitting point and the decrease in squared errors across all trees (12).

c. Partial dependence profiles (PDPs): PDPs provide visual feedback for the interpretation of any black box model by showing the influence of different features or feature subsets. It also shows the marginal effect of a variable for the average prediction (5). The impact of a variable can be determined by observing the change in the average response (12). PDPs do not take into consideration all possible feature interactions. Therefore, they can provide limited accurate information about the model. Despite this, they frequently provide helpful information, which substantially aids in understanding black box models, particularly when most of these interactions are low. They can be used for improving

ML models, comparing different models, and evaluating model performance (5). PDP plots were created for the four most important parameters (glucose, age, waist circumference, and mean corpuscular volume (MCV) to investigate and compare the prediction patterns of the two models.

Local exploration (single row prediction): In contrast to global exploration, local exploration methods assist in comprehending how a model generates a prediction for an individual data point. In the study, cases with the highest model prediction score and incorrect predictions were selected for local exploration. The following local explorations were used:

a. Breakdown plots for additive attributions: Breakdown explainability is a method for understanding the contribution of each feature to a specific prediction made by a ML model. It breaks down the prediction into the contributions of individual features and shows how each feature contributes to the final prediction. This type of explainability is particularly useful for understanding how a model arrived at a specific prediction and can help identify any biases or errors in the model (5, 18).

b. SHAP: This method assigns a unique importance value to each feature, indicating the contribution of that feature toward the model's output for a particular prediction. It is based on the game theory and aims to improve interpretability by calculating the significance of each feature for individual predictions. The objective of SHAP is to clarify the prediction of a specific data point x by calculating the impact of each feature on the prediction. The method utilizes certain visualization techniques to display how the predictors affect the predicted values. It also allows for the identification of feature interactions and provides global and local explanations (17, 18).

RESULTS

The demographic features of patients and clinical laboratory result summaries are given in **Table 1**. As shown in this table, the records had wide spectrum of laboratory results. Additionally, there was a distinctive class imbalance between A1c classes and sex.

Table 2 summarizes the prediction errors for DRF and GBM models. There was no misclassification for the DRF ML model on the training set. However, prediabetes and diabetes prediction error rates were lower for the GBM model on the test set.

Table 3 provides the performance metrics for both models. When the performance metrics were evaluated, both models had similar AUC and AUCPR metrics. However, DRF had a lower mean per class error for the test set, while DRF was lower for the test set.

In **Figure 1**, PDF plots were given for the top important features for both models, which provides information related to all data sets. Figures 2 and 3 show the breakdown and SHAP plots, respectively, for four selected test set records. These plots cover three correct classifications and one misclassification for the DRF model. All model details, including the hyperparameters, are provided in the Supplementary Material.

DISCUSSION

The study aimed to compare the performance of different model-agnostic explanation methods using two different ML models created for A1c prediction. The results showed that model explanation methods provide important contributions for both evaluating models internally and comparing different models. Moreover, it was found that using a combination of local and global explanation models is more effective for explaining and comparing models than using a single

Table 2. Multinominal classification confusion matrix for the developed ML models

Model	Class	Training Set				Test Set			
		Normal	PreD.	DM	Error	Normal	PreD.	DM	Error
DRF	Normal	1329	0	0	0.0%	548	21	0	3.6%
	PreD	0	551	0	0.0%	178	52	7	78%
	DM	0	0	244	0.0%	36	17	53	5%
GBM	Normal	1259	67	3	5.3%	518	50	1	8.9%
	PreD	187	356	8	35.4%	117	101	19	57%
	DM	6	18	220	9.9%	5	23	78	26%

DRF: Distributed Random Forest, GBM: Gradient Boosting Machine, PreD: Prediabetes

Table 3. Calculated performance metrics for the developed ML models

Model	Data Set	Mean per Class Error	AUC	AUC Precision Recall	Accuracy	Macro F1
DRF	Training	0.00	1.00	1.00	1.00	1.00
	Test	0.44	0.85	0.84	0.72	0.63
GBM	Training	0.17	0.96	0.96	0.86	0.85
	Test	0.31	0.88	0.86	0.76	0.71

AUC: Area under curve, DRF: Distributed Random Forest, GBM: Gradient Boosting Machine

explanation model. However, there are some limitations related to these tools, such as limited interpretability and the required computation power. Although these tools helped the end user (e.g., clinicians) to understand some predictions, they only provided a general idea.

The first finding that model explanation techniques contribute to prediction explanations is supported by the conclusions derived from PDP plots. When examining the PDP graphics in **Figure 2**, the relationship between glucose levels and A1c is clearly visible. The known relationship between glucose and A1c is that estimated average glucose (eAG)= 28.7×A1c(%)– 46.7 can be easily observed with both models (19). In both the DRF and GDM models, the predictions for A1c quickly change according to the blood glucose levels. The effect of age can also be seen in a similar way in these plots. It has also been mentioned in the literature that PDP graphics can be easily interpreted by field experts because they provide explanation by simplification (3). Another use of PDP is to compare the effect of features for different models, as shown in **Figure 2**. The effect of waist circumference, which is more meaningful for the DRF model, can be observed in **Figure 2**. Relation between DM and waist circumference was also reported by Feller et al. (20). However, when it comes to the interaction between multiple parameters, the power of these graphics decreases. The assumption of independence is the biggest issue with PDP (5).

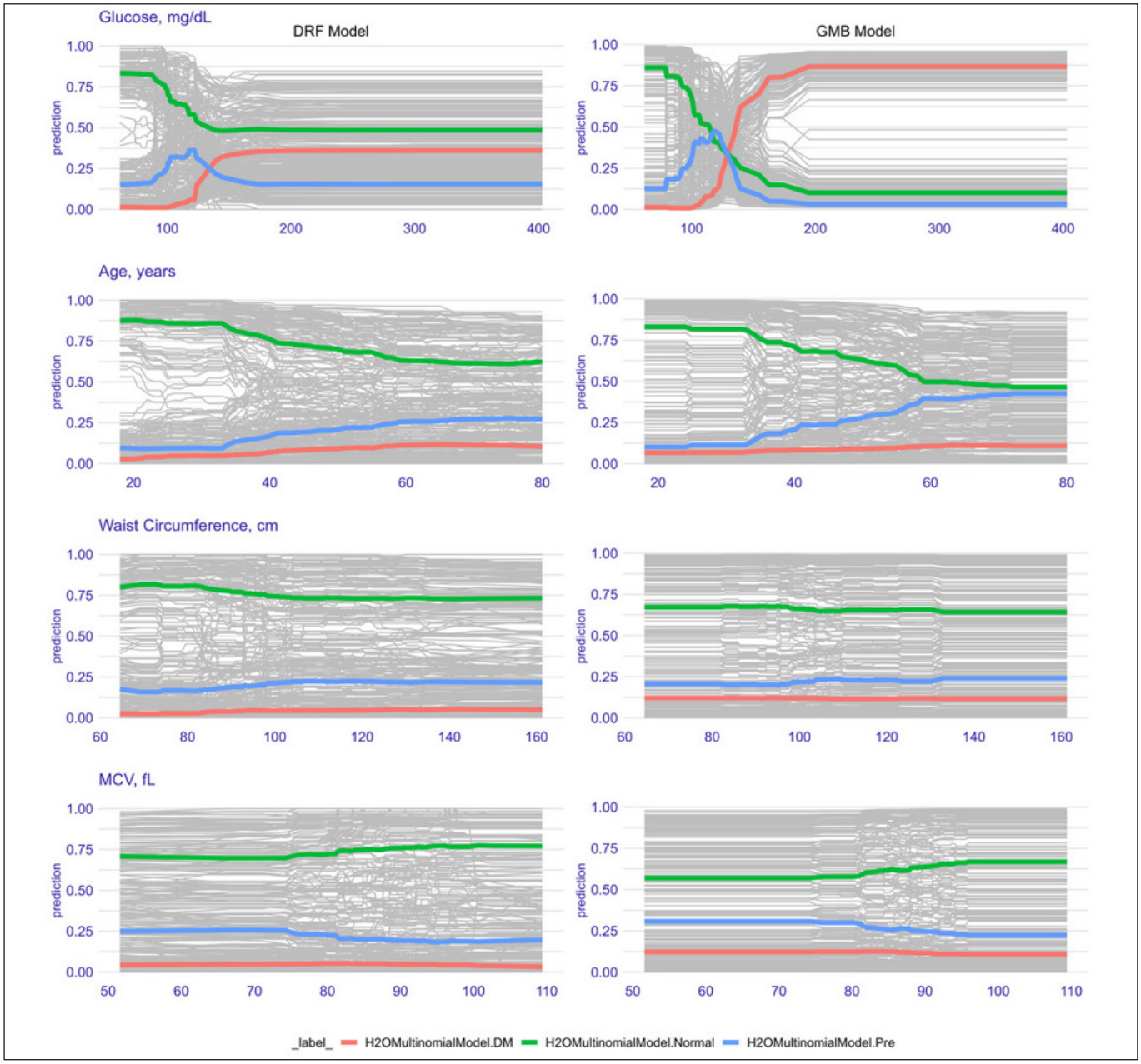

Figure 1. Partial dependence plots for the four most important features of the DRF and GMB models.
The Partial dependence lines for normal prediction classes are green, pre-diabetes prediction classes are blue, diabetes mellitus prediction classes are yellow, and ceteris-paribus profiles are gray lines.
DRF: Distributed Random Forest, GBM: Gradient Boosting Machine.

Local explanation techniques explain specific predictions, which provide details about the inner workings of the model instead of using all model data. Therefore, they can be also used for understanding the model's decision-making process. Breakdown plots in **Figure 2b** precisely show each feature's contribution to the final prediction. The relationship between serum glucose and age, which can also be observed by global methods, on normal, prediabetes, and diabetes predictions are clearly shown case by case. Especially for unexpected predictions, local explanation methods reveal the cause of the model's failed reasoning, which makes it possible to improve the model (18). For example, **Figure 2d** shows that an individual with a diabetic A1c level has been incorrectly classified as normal. This situation can be caused by majority case drift caused by the number of records with normal A1c levels during the development of the model, which presents a 0.716 intercept score in **Figure 2**. When detecting the issue, different data pre-processes could be considered for imbalanced data to improve the model. SHAP plots also allow for the identification of feature interactions and provide global and local explanations. SHAP values can also be used for variable importance by calculating them for the entire dataset (21).

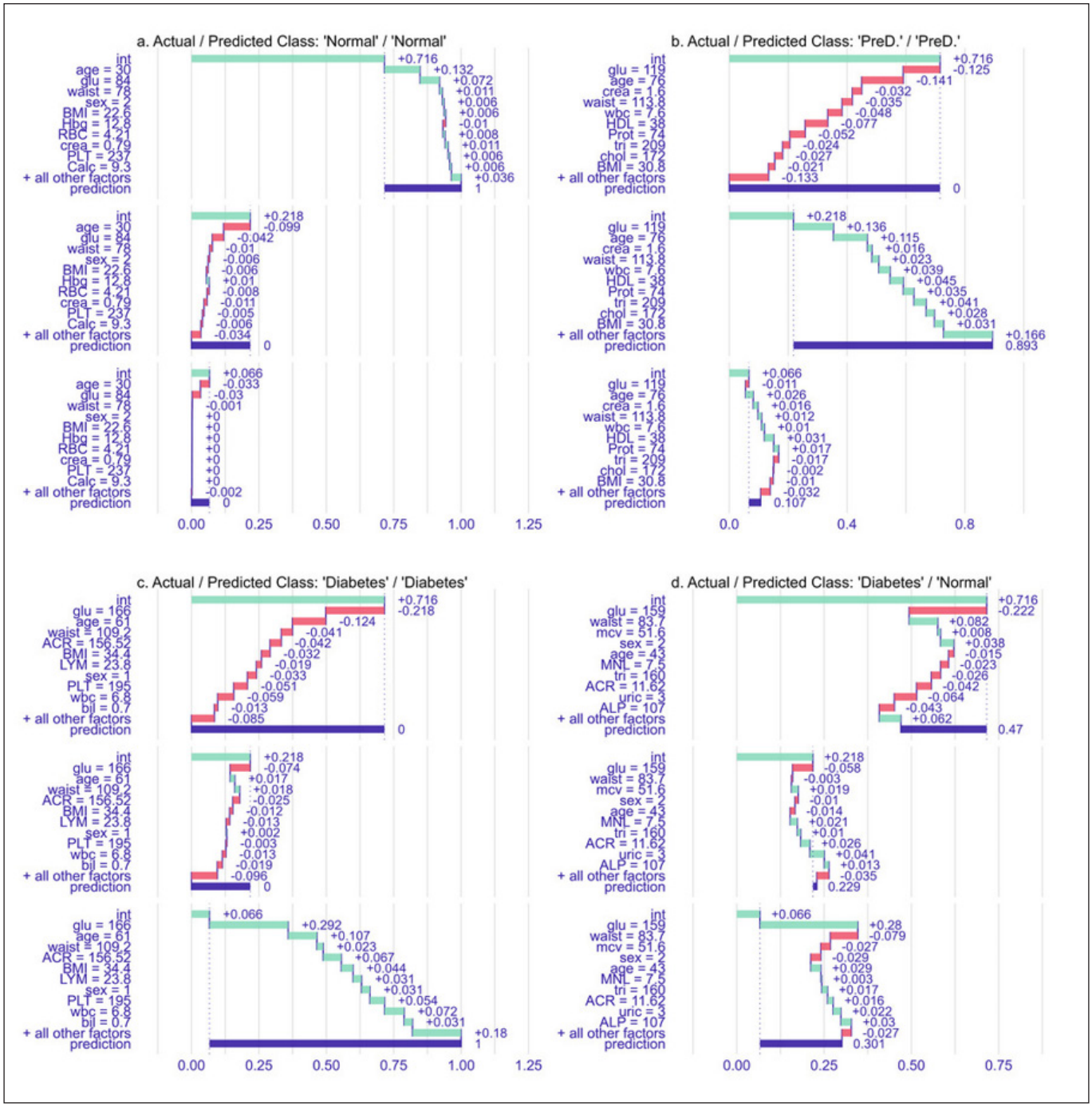

Figure 2. Breakdown plots for additive attributions of Distributed Random Forest model for the test set.
The green and red bars indicate positive and negative changes in the mean predictions, respectively. The blue bar shows the prediction for the instance of interest.

Another important finding was for the DRF ML model, while there was no classification error for the training set, there were classification errors in the test set as expected (**Table 2**). Moreover, the training set had better performance metrics compared to test set (**Table 3**). These results suggest overfitting for the DRF model. Therefore, limiting the depth of individual trees, increasing the number of trees in the forest, and using techniques such as bagging and feature subsetting can help to prevent overfitting (12). Additionally, in both data sets classification between normal and prediabetic patients was more distinctive than prediabetic and diabetic classification (**Table 2**). Interestingly this finding was similar with clinical setting and this distinction is considered as challenging. Prediabetes is a condition that exists on a continuum between normal blood glucose levels and diabetes. And is characterized by higher-than-normal blood glucose levels but not high enough to be classified as diabetes. The borderline nature of prediabetes means that there can be significant overlap between normal and pre-diabetic patients in terms of their blood glucose levels, making it challenging to accurately separate the two groups (8, 19).

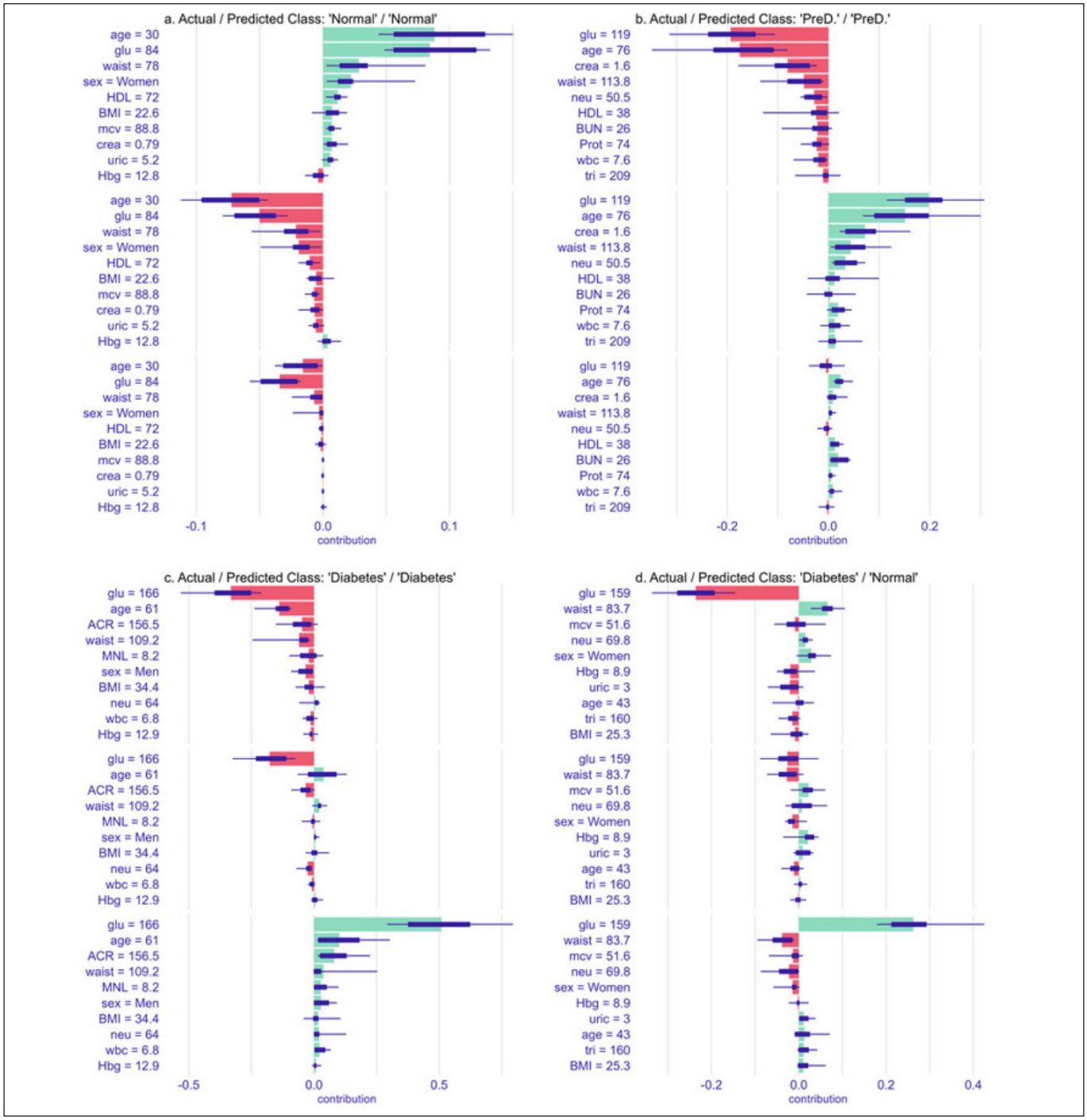

Figure 3. Plots of Distributed Random Forest model Shapley values for the test set.
Red and green bars present negative and positive Shapley values, respectively.

In the study, model-agnostic methods were applied for xAI. Model-agnostic explanations are considered consistent across different models, which means they can be easily used for multiple model comparisons. Additionally, their model-independent nature provides developers with more options for selecting models during development. An alternative to model-agnostic interpretation is using only interpretable models, but this may result in reduced performance and limits the choice of models that can be used (5).

Different methods have been proposed to assess and quantify the quality of explanations generated by xAI systems, as shown in this study. Nonetheless, there is currently no widely accepted standard for determining whether an xAI system is more understandable to a user than a non-xAI system. Some methods rely on subjective evaluations such as surveys to gauge user satisfaction with the explanations. Other methods are more objective, such as determining whether the explanations improve the user's decision-making performance (4). Instead of providing specific, valid justifications for a model's predictions, it is more accurate to view explainability techniques as overall explanations of how a model operates (3). As in the current study, these tools are still far from interpreting results in a clinical context.

This research has several limitations. First, the use of glucose for A1c prediction can be considered a bias. However, the glucose parameter was specifically selected for the demonstration of local and global explanations of strong and weak features. Second, preprocessing for imbalance classes for both diabetes classes and sex could also increase model effectiveness. However, since the aim of the study is not to create models with the best performance but to evaluate the effectiveness of explanation tools, simpler methods were preferred while creating models. Additionally, during training, the "imbalanced_classes" option was activated in the AutoML tool. The final limitation refers to the included ML model types, as only tree-based models were included in the study. However, using different models such as deep learning can provide different perspectives for model explainability.

CONCLUSION

The current study assessed the importance of transparent ML methods in understanding complex models and how it promotes the integration of AI in healthcare. Results showed that despite the limitations of current explainability methods, especially for the clinical approach, both global and local explanation models provide an insight into model evaluation, and they can be used to improve or compare models.

Abbreviations

A1c: hemoglobin A1c, **AUC**: area under the receiver operating characteristic curve, **AUCPR**: area under the precision recall curve, **DRF**: default random forests, **eAG**: estimated average glucose **ML**: machine-learning, **xAI**: explainable artificial intelligence, , **GBM**: Gradient boosting machine, **MCV**: mean corpuscular volume, **PDPs**: Partial Dependence Profiles, **SHAP**: Shapley additive explanations, Variable Importance (VIP)

ETHICAL DECLARATIONS

Ethics Committee Approval: Not applicable. An open-source dataset was utilized for the study.

Referee Evaluation Process: Externally peer-reviewed.

Conflict of Interest Statement: The author has no conflicts of interest to declare.

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

Author Contributions: The author declares that he has responsible for the design, execution, and analysis of the paper and that he has approved the final version.

REFERENCES

1. Haymond S, McCudden C. Rise of the Machines: artificial ıntelligence and the clinical laboratory. J Appl Lab Med 2021; 6: 1640–54.
2. Habehh H, Gohel S. Machine learning in healthcare. Curr Genomics. 2021; 22: 291-300.
3. Zhang Y, Weng Y, Lund J. Applications of explainable artificial ıntelligence in diagnosis and surgery. Diagnostics 2022; 12: 237.
4. Arbelaez Ossa L Starke G, Lorenzini G, Vogt JE, Shaw DM, Elger BS. Re-focusing explainability in medicine. Digit Heal 2022; 8: 205520762210744.
5. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable ai: a review of machine learning interpretability methods. Entropy 2020; 23: 18.
6. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Heal 2021; 3: e745-50.
7. Langs HG, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. Wiley Interdiscip Rev Data Min Knowl Discov 2019; 9: e1312.
8. Sherwani SI, Khan HA, Ekhzaimy A, Masood A, Sakharkar MK. Significance of HbA1c test in diagnosis and prognosis of diabetic patients. Biomark Insights 2016; 11: 95–104.
9. National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. Hyattsville, MD; U.S. Department of Health and Human Services, Centers for Disease Control and Prevention. [December 2022][https://www.cdc.gov/nchs/nhanes/ index.htm]
10. American Diabetes Association Professional Practice Committee (2022). 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes. Diabetes care 2022; 45: 17-38.
11. R Core Team R. A language and environment for statistical computing. https://www.r-project.org/ [December 2022].
12. Ledell E, Poirier S. H_2O AutoML: scalable automatic machine learning [Internet]. AutoML Org; 2020. [https://H_2O.ai/platform/H_2O-automl/] [December 2022]

13. Biecek P. DALEX: Explainers for Complex Predictive Models in R. J Mach Learn Res 2018; 19: 1-5.
14. Kursa MB, Rudnicki WR. Feature selection with the boruta package. J Stat Software 2010;36:1–13.
15. Hand DJ, Till RJ. A simple generalisation of the area under the ROC curve for multiple class classification problems. Mach Learn 2001; 45: 171-86.
16. Wei P, Lu Z, Song J. Variable importance analysis: A comprehensive review. Reliab Eng Syst Saf 2015; 142: 399-432.
17. Antwarg L, Miller RM, Shapira B, Rokach L. Explaining anomalies detected by autoencoders using Shapley Additive Explanations. Expert Syst Appl 2021; 186: 115736.
18. Staniak M, Biecek P. Explanations of model predictions with live and breakDown packages. R Journal 2019; 10: 395-409.
19. Nathan DM, Kuenen J, Borg R, Zheng H, Schoenfeld D, Heine RJ. Translating the A1C assay into estimated average glucose values. Diabetes Care 2008; 31: 1473-8.
20. Feller S, Boeing H, Pischon T. Body mass index waist circumference and the risk of type 2 diabetes mellitus: implications for routine clinical practice. Dtsch Arztebl Int 2010: 470-6.
21. Shimizu T, Suda K, Maki S, et al. Efficacy of a machine learning-based approach in predicting neurological prognosis of cervical spinal cord injury patients following urgent surgery within 24 h after injury. J Clin Neurosci Off J Neurosurg Soc Australas 2023; 107: 150-6.