ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2.
Available from http://www.i6doc.com/en/.

Influence of Data Characteristics on Machine Learning Classification Performance and Stability of SHapley Additive exPlanations

Anusha Ihalapathirana[1], Gunjan Chandra[1], Piia Lavikainen[2], Pekka Siirtola[1], Satu Tamminen[1], Nirzor Talukder[1], Janne Martikainen[2] and Juha Röning[1] *

1- Biomimetics and Intelligent Systems Group, University of Oulu, Finland
2- School of Pharmacy, University of Eastern Finland, Kuopio, Finland

Abstract. This study explores the effects of different data sizes and data imbalance on model performance and the stability of SHapley Additive exPlanations (SHAP). The study utilizes a Type 2 diabetes (T2D) dataset to train three machine learning (ML) models: linear discriminant analysis, XGBoost, and a neural network. It shows that adjusting the background dataset size leads to variations in the SHAP values, with decreased variance observed in larger and balanced datasets. Furthermore, the study highlights that the data characteristics leading to high model performance may not always produce reliable and stable SHAP explanations.

1 Introduction

Artificial intelligence (AI) models greatly benefit from a large amount of data during the training process. Despite the abundance of big data in various fields, many real-world medical datasets suffer greatly from an imbalanced class distribution, particularly when it comes to rare diseases. This significantly impacts rare event detection, as most classifiers implicitly assume a balanced class distribution and aim to maximize overall accuracy, leading them to favor the majority class.

In machine learning applications, understanding the reasoning behind model decisions is important for end-users across various domains, particularly in fields like medicine. Although achieving the optimal model for medical applications may not always be feasible, prioritizing transparency ensures that users comprehend the model's functioning and enhances its reliability and trustworthiness. Post-hoc explanation methods are necessary to interpret certain ML models, which can be too complex even for experts to interpret. SHapley Additive exPlanation [1] is one such post-hoc explanation method that requires a background dataset when interpreting ML models. The background dataset consists of representative data samples used as a reference to compute the expected values of the model outputs. Therefore, identifying the influence of the background dataset on the quality and reliability of SHAP explanations is important to ensure the model's trustworthiness.

*This work was funded by HTx project. The HTx project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement № 825162.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

The existing literature extensively examines how the size of the data and the class imbalance affect model performance [2, 3, 4]. However, none of these studies considered the effects of data characteristics on the results of the explanation methods. Recently, Yuan et al. [5] investigated the effect of background data size on the stability of SHAP in deep learning models. They also investigated the effects of data imbalance on SHAP explanations for deep learning models [6]. However, both studies were conducted using a single artificial neural network model, without consideration of model performance. Furthermore, they did not simultaneously consider data size and class imbalance of the background data in their analyses. Considering both factors together would provide a more comprehensive understanding of their combined influence on SHAP explanations. Therefore, this article furthers the investigation to explore the impact of data size and data imbalance on the performance and the SHAP explanation of machine learning models using Type 2 diabetes (T2D) data. It aims to understand whether high model performance implies corresponding reliability and stability in the model explanation.

2 Methodology

This study incorporates and expands on the research findings and methodology presented by Lavikainen et al. [7]. The study [7] was carried out with the T2D dataset to identify long-term glycemic control clusters of patients based on their six-year glycated hemoglobin (HbA1c) trajectories. The binary classification was then performed to predict the trajectory membership, 'stable+adequate' and 'inadequate' classes. Furthermore, the study used three types of predictors: clinical-, treatment-, and socioeconomic-related. In this study, we focus only on models developed with clinical-related predictors.

The dataset used in this study is the North Karelia Wellbeing Services County electronic health register data, the same dataset used by [7]. In our analysis, we use three classification models, including the final models selected by [7] - linear discriminant analysis (LDA) and neural network (NN) - along with the addition of XGBoost (XGB). All models were trained using the same set of hyperparameters across all scenarios, with each model using five features, HbA1c 1 year before (mmol/mol), HbA1c 2 years before (mmol/mol), fasting plasma glucose (mmol/l), T2D duration (years), and Other cardiac diseases (other diseases of the heart and pulmonary circulation).

The data preprocessing steps, methods, and parameters are the same as [7], except that in our study, all the duplicate entries have been removed. After preprocessing, the dataset contained 7601 samples and we used random sampling to partition 500 samples from the preprocessed dataset to use as a test dataset. The performance of various model configurations in this study was evaluated using this test dataset. The remaining data was used as the training data during the model training phase and as background data for SHAP value calculations. Each model was trained with a different percentage of the dataset, ranging from 10% to 100% of the original training data, and we used the same training dataset

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

as the background data. Table 1 displays the number of samples for each data size. All the models were trained with the same five predictors mentioned above and validated using 10-fold cross-validation. The performance of all models was evaluated using balanced accuracy (BA), F1 score, and area under the ROC curve (AUC).

	10%		**25%**		**50%**		**75%**		**100%**	
	Train	Val	Train	Val	Train	Val	Train	Val	Train	Val
C0	543	109	1363	279	2717	579	4064	869	5401	1162
C1	41	7	98	18	207	42	322	65	447	91

Table 1: Number of training and validation samples in each data size. The test set consists of a fixed 500 samples across all data sizes.

C0: Adequate Class, C1: Inadequate Class

First, the study was conducted using imbalanced data. Subsequently, the steps were repeated using balanced data, achieved by applying the Synthetic Minority Over-sampling Technique (SMOTE) [8] to oversample and balance the dataset in each scenario. In the next step, we assess SHAP explanations using the mean absolute SHAP values across splits in 10-fold cross-validation for each feature, using *shap.Explainer*, under various background data sizes and class distributions[1].

3 Results and Discussion

Table 2 presents the performance of all models trained with different configurations. LDA achieved its highest balanced accuracy of 0.882 using 25% of balanced training data, indicating a significant increase compared to the models trained with imbalanced data. Increasing the amount of training data shows a significant impact on the performance of the NN model. However, there may be a point of saturation, beyond which additional training data does not significantly improve model performance. Moreover, different ML models require different amounts of training data for optimal performance. In our study, the impact of data size on model performance is minimal irrespective of the model, with noticeable effects primarily observed in NN models. In this study, the results indicate that balancing the class-wise distribution of samples using SMOTE leads to an improvement in the balanced accuracy evaluation metrics, but it did not always result in an improvement in the F1 score and AUC.

Figure 1 shows the variations in the mean absolute SHAP values of the features across cross-validation splits within the validation set, for both balanced and imbalanced data. The LDA model exhibits higher mean absolute SHAP values, while the NN model demonstrates the lowest, indicating less certainty in its interpretation for this dataset. Furthermore, the results show that models built

[1]The source code accessible on GitHub: `https://github.com/anushaihalapathirana/data_characteristics_shap`

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

	Model		Training Data size in Percentage (%)				
			10	**25**	**50**	**75**	**100**
Imbal. data	LDA	BA	**0.772**	0.758	0.759	0.732	0.733
		F1	**0.734**	0.728	0.731	0.715	0.718
		AUC	**0.914**	0.914	0.913	0.913	0.913
	XGB	BA	0.643	0.682	**0.691**	0.678	0.647
		F1	0.666	**0.741**	0.730	0.721	0.682
		AUC	0.907	**0.910**	0.899	0.885	0.900
	NN	BA	0.683	0.690	0.700	0.698	**0.713**
		F1	0.699	0.725	0.720	0.712	**0.728**
		AUC	0.887	0.911	**0.911**	0.907	0.908
Bal. data	LDA	BA	0.881	**0.882**	0.860	0.857	0.858
		F1	0.691	0.668	**0.695**	0.688	0.690
		AUC	0.913	0.913	**0.915**	0.914	0.914
	XGB	BA	0.783	**0.799**	0.794	0.791	0.791
		F1	0.700	**0.714**	0.699	0.691	0.691
		AUC	0.900	0.892	**0.902**	0.863	0.888
	NN	BA	0.736	0.750	0.796	0.811	**0.813**
		F1	0.641	0.672	0.678	**0.685**	0.645
		AUC	0.864	0.855	**0.895**	0.892	0.886

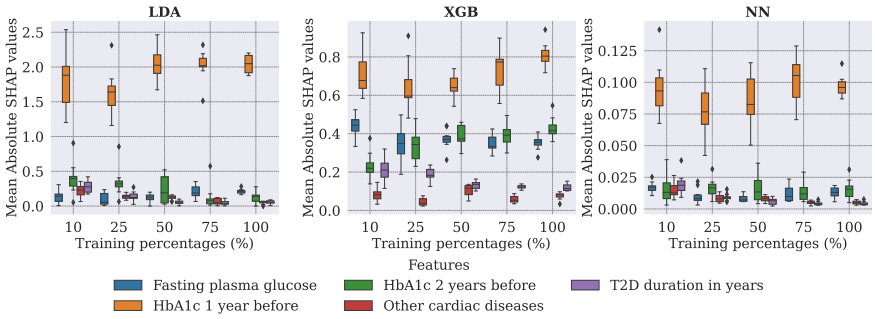
Table 2: Performance of models on the test dataset.

Imbal.: Imbalanced, Bal.: Balanced, F1: F1 score.

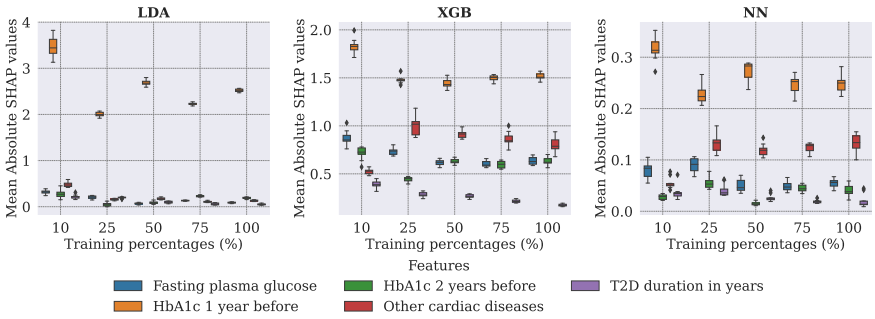
on balanced data tend to have higher mean absolute SHAP values compared to those with imbalanced data. Additionally, the findings reveal that SHAP is more reliable in ranking the most important feature; however, the ranking of other features tends to fluctuate with different data sizes. Moreover, larger background datasets lead to more stable SHAP values. It is particularly noticeable in the LDA model, where the SHAP values show less variation with increased data size. For example, the variances of *other cardiac diseases* feature in the LDA model, using balanced data across data sizes ranging from 100% to 10%, are as follows: 0.0028, 0.0003, 0.0004, 0.0002, and 0.0001. It is important to note that even with more extensive datasets, the issue of imbalanced data can result in potential inconsistencies in feature importance rankings.

We used Uniform Manifold Approximation and Projection (UMAP) [9] to analyze the correlation between the mean absolute SHAP values of features in cross-validation splits and background data sizes (Figure 2). UMAP demonstrated more effective clustering of SHAP values across different data sizes when using balanced background data as opposed to imbalanced data. Furthermore, even when using balanced background data, linear models tend to result in denser clusters, while complex models tend to result in sparser clusters. This highlights the impact of model complexity, as well as the size and balance of the background data, on SHAP explanations. In all cases, UMAP demonstrated that the size of the background data significantly influences SHAP values.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.



(a) Mean Absolute SHAP Values Calculated with Imbalanced Background Data



(b) Mean Absolute SHAP Values Calculated with Balanced Background Data

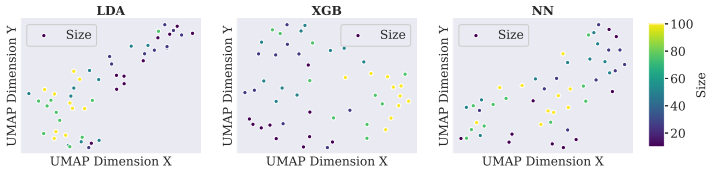
Fig. 1: Comparative Analysis of Mean Absolute SHAP Values for Features Across 10-Fold Cross-Validation Splits.

4 Conclusion

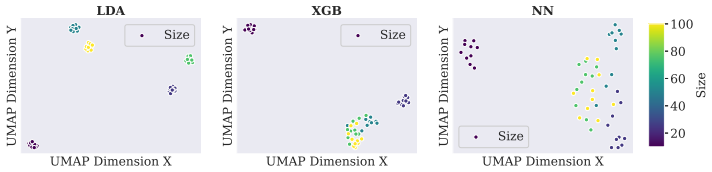
This study investigated the effect of data size and imbalance on the performance and SHAP explanation of three different machine learning models. Our study highlights the significance of explaining both the rationale behind model decisions and the potential consequences of machine learning models to identify how reliable the models are, enhancing trust and efficacy in decision-making processes. The study revealed that the data size leading to high model performance does not always result in correspondingly reliable and stable SHAP explanations. It is important to note that, to obtain reliable and stable SHAP explanations, researchers should refrain from using excessively small background data sizes.

A limitation of our study was that it focused on a single dataset and examined only the influence of data size and imbalance on the stability of SHAP values. Other factors, like domain-specific context and feature interactions, also contribute to understanding the model. We plan to extend our work by investigating various datasets alongside diverse data balancing techniques and data augmentation methods, as well as evaluating different validation metrics and conducting computational cost analysis, to enhance the study analysis.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.



(a) Mean Absolute SHAP Values Calculated with Imbalanced Background Data



(b) Mean Absolute SHAP Values Calculated with Balanced Background Data

Fig. 2: UMAP Visualization of Mean Absolute SHAP Values for Features Across 10-Fold Cross-Validation Splits.

References

[1] S. Lundberg and S-I. Lee, A Unified Approach to Interpreting Model Predictions, arXiv:1705.07874, arXiv preprint, 2017.

[2] C. A. Ramezan and T. A. Warner and A. E. Maxwell and B. S. Price, Effects of Training Set Size on Supervised Machine-Learning Land-Cover Classification of Large-Area High-Resolution Remotely Sensed Data, Remote Sensing, 13:368, Multidisciplinary Digital Publishing Institute, 2021.

[3] P. T. Noi and M. Kappas, Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery, Sensors, 18:18, Multidisciplinary Digital Publishing Institute, 2018.

[4] A. Bailly and C. Blanc and É. Francis and T. Guillotin and F. Jamal and B. Wakim and P. Roy, Effects of dataset size and interactions on the prediction performance of logistic regression and deep learning models, Computer Methods and Programs in Biomedicine, 213:106504, 2022.

[5] H. Yuan, and M. Liu and M. Krauthammer and L. Kang and C. Miao and Y. Wu, An empirical study of the effect of background data size on the stability of SHapley Additive exPlanations (SHAP) for deep learning models. International Conference on Learning Representations (ICLR 2023), 2023.

[6] M. Liu and Y. Ning and H. Yuan and M. Ong and N. Liu, Balanced background and explanation data are needed in explaining deep learning models with SHAP: An empirical study on clinical decision making, arXiv:2206.04050, arXiv preprint, 2022.

[7] P. Lavikainen and G. Chandra and P. Siirtola and S. Tamminen and A. T. Ihalapathirana and J. Röning and T. Laatikainen and J. Martikainen, Data-Driven Identification of Long-Term Glycemia Clusters and Their Individualized Predictors in Finnish Patients with Type 2 Diabetes, Clinical Epidemiology, 31:13-29, 2023.

[8] N. Chawla and K. Bowyer and L. Hall and W. Kegelmeyer, SMOTE: Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research, 16:321-357, 2002.

[9] L. McInnes and J. Healy and J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, arXiv:1802.03426, arXiv preprint, 2020.