Pairwise Difference Filter (PDF): An Interpretable Preprocessing Method for Medical and Beyond

Daniel Pordeus[1], **Weslley Lioba Caldas**[1], **João Paulo do Vale Madeiro**[1]

[1] Computer Science Department – Federal University of Ceará (UFC)
Fortaleza – CE – Brasil

pordeus@alu.ufc.br, weslleylc@gmail.com, jpaulo.vale@dc.ufc.br

Abstract. *Interpretability is a critical requirement for machine learning models in healthcare, as clinicians need to understand how the input features are processed to make informed decisions about patient care. Traditional preprocessing methods often fail to capture subtle differences between patient groups, particularly in datasets with overlapping or highly similar classes. To address these challenges, we propose* ***Pairwise Difference Filter (PDF)****, a novel preprocessing method that leverages pairwise differences between samples of opposite classes to identify the most influential features. PDF focuses on pairs of patients with the smallest overall differences but significant differences in specific features, enabling the identification of clinically meaningful biomarkers. By enhancing the interpretability of machine learning models, PDF supports medical decision-making and improves the transparency of predictive models in healthcare. Experimental results with three different on a COVID-19 severity classification dataset, MUSIC (a dataset for predicting outcomes in patients with several degrees of heart failure) and Wine Toy Dataset demonstrate that PDF achieves competitive performance while providing interpretable feature rankings that align with clinical knowledge.*

1. Introduction

Interpretability is a critical requirement for machine learning models in healthcare, such tht clinicians can potencially apply these models to make informed and trustworthy decisions about patient care. The importance of interpretability in the medical field has been widely emphasized, particularly in ensuring that predictions are not only accurate but also understandable and actionable for healthcare professionals [Hakkoum et al. 2022]. [Lisboa et al. 2023] discuss the evolution of interpretable and explainable machine learning models, highlighting their role in modern data-driven decision making. They argue that interpretability is essential for building trust in machine learning systems, especially when these systems are used to support critical decisions. Traditional preprocessing methods, such as those based on tree-based importance scores or permutation importance, often fail to capture subtle differences between patient groups, particularly in datasets with overlapping or highly similar classes [Jović et al. 2015]. This limitation can hinder the interpretability of machine learning models and reduce their utility in clinical practice, where understanding the underlying factors driving predictions is essential for decision-making [Hakkoum et al. 2022].

To address these challenges, we propose the Pairwise Difference Filter (PDF), a novel preprocessing method that leverages pairwise differences between samples of opposite classes. By focusing on pairs of patients with the smallest overall differences but

significant differences in specific features, PDF identifies clinically meaningful biomarkers that enhance the interpretability of machine learning models. This approach is particularly useful for medical applications, where interpretability is not merely a technical requirement but a fundamental necessity for ensuring trust, transparency, and adoption in clinical workflows [Hakkoum et al. 2022]. By bridging the gap between model performance and interpretability, PDF offers a promising solution to improve the utility of machine learning in healthcare, enabling clinicians to make more informed and reliable decisions.

This paper is organized as follows. Section 2 presents other research related to the topic of this work. Section 3 explains the proposed preprocessing method step by step through an example. Section 4 describes the datasets that were used for performance comparison. Section 5 presents the results of the performance comparison. Finally, Section 6 discusses the results and draws conclusions.

2. Related Work

In this section we will present some examples of filtering and feature selection methods that share some kind of conceptual similarity with the **Pairwise Difference Filter (PDF)**. In some cases we will mention the difference or the complement that the PDF has with the mentioned method.

Introduced by [Kira and Rendell 1992], the **Relief** algorithm is a non-parametric method that assigns weights to features based on their ability to distinguish between nearby instances (*near-hit* and *near-miss*). Its core principle is that relevant features should have similar values for observations in the same class and dissimilar values for different classes. [Kononenko 1994] extended the method for multi-class problems (**ReliefF**), incorporating probabilistic calculations and robustness to missing data. It is effective for non-linear relationships between features and target.

SFS is a wrapper-based approach that iteratively selects feature subsets by adding (*forward selection*) or removing (*backward elimination*) attributes based on model performance. Formalized by [Kohavi and John 1997] as a greedy algorithm, SFS optimizes feature subsets directly, but with a high computational cost for large datasets.

[Caldas 2024] propose a feature selection method based on perfect bipartite matching, which shares some conceptual similarities with PDF. However, their approach focuses on optimizing a global objective, whereas PDF emphasizes local discriminative patterns. This makes PDF particularly suitable for datasets with overlapping or highly similar classes.

[Dhurandhar et al. 2018] introduce a method to generate contrastive explanations using pertinent negatives. Their approach answers the question: *"Why did the model predict X instead of Y?"* by identifying features that, if changed, would lead to a different prediction. Although PDF does not generate counterfactuals, it shares a similar focus on understanding the differences between classes.

[Van Looveren and Klaise 2021] propose a method to generate interpretable counterfactual explanations guided by prototypes. Their approach ensures that counterfactuals are realistic and actionable, making them useful for applications such as healthcare and finance. PDF complements this work by providing a preprocessing method that enhances

the interpretability of predictive models without relying on synthetic data.

3. Pairwise Difference Filter Method

In this work, we propose a novel preprocessing method called **Pairwise Difference Filter (PDF)**. The inspiration for creating PDF came from studying counterfactual-based explainability methods such as [Ribeiro et al. 2016] and [Van Looveren and Klaise 2021]. PDF is designed to identify the most influential features by analyzing pairwise differences between samples of opposite classes. The method focuses on pairs of patients from opposite classes with the smallest overall differences but significant differences in specific features, enabling the identification of clinically meaningful biomarkers. This approach enhances the interpretability of the data, emphasizing, for example, which data contributed most to differentiating two different clinical cases. This approach in itself is already interesting to assist in medical analysis. The application of machine learning models corroborates the validation of the method, reinforcing that it can be useful for medical applications where understanding the underlying factors that drive predictions is essential for clinical decision-making. To facilitate understanding of the method, a fictitious example will be presented for didactic purposes.

3.1. Step 1: Data Normalization

The dataset is preprocessed to normalize features and ensure consistency. We will use a fictitious example, presented in Table 1, so that the explanation of the method is as didactic as possible. Each sample is represented by features such as age, weight, height, and body mass index (BMI), along with the target variable (e.g., cause of death).

Age	Weight (kg)	Height (cm)	BMI (kg/m^2)	Cause of Death
0.71831	0.591398	0.746032	0.393939	1
0.732394	0.784946	0.809524	0.524242	0
0.760563	0.526882	0.666667	0.384848	0
0.690141	0.645161	0.571429	0.578788	1
0.422535	0.344086	0.761905	0.154545	1

Table 1. MinMaxScaled data for PDF analysis.

3.2. Step 2: Pairwise Difference Calculation

Let x_j^Z be the feature j of class Z. For each pair of patients from opposite classes (e.g., class 0 and class 1), we compute the sum of absolute differences across all features. This value, called **DIFF**, represents the overall dissimilarity between the two patients:

$$DIFF = \sum_{i=0}^{n} |(x_i^0 - x_i^1)|$$

Table 2 shows an example of the calculated differences.

Age	Weight (kg)	Height (cm)	BMI (kg/m^2)	DIFF
0.014084	0.193548	0.063492	0.130303	0.401427

Table 2. Pairwise differences between samples.

3.3. Step 3: Sorting and Selection

We sort all pairs by **DIFF** in ascending order and select pairs where **DIFF** falls within the interval $[min, min + std]$, where min is the smallest **DIFF** value and std is the standard deviation of **DIFF**. This ensures that we focus on pairs of patients that are similar overall but may have significant differences in specific features. Table 3 shows the result for this example of pairs sorted ascending by **DIFF** values. The values in columns **Class 0** and **Class 1** are the indexes of each sample in the dataset. In other words, in the first row, we have that sample 46 is from class 0, and is confronted with sample 479, which is from class 1.

Class 0	Class 1	DIFF
46	479	4.4708
84	58	4.7002
226	134	4.7516
332	419	5.3119
762	753	5.3787

Table 3. Ranked features based on pairwise differences.

3.4. Step 4: Feature Ranking

The final step involves selecting the top-ranked features as clinically meaningful biomarkers. This step consists of checking which features were the most misbehaved - features whose values had the largest delta - among the pairs selected in Step 3. So, we calculate the sum of the differences of each feature, ranking in descending order by this computed value. With this ranking in hand, we can choose how much data to use to train the model. The option we were adopting was to select the data from the range [max, max+std], where max is the maximum value in the ranking and std is the standard deviation of the ranking values, but when the original number of features is small, we can choose to select 25 or 50% of the data. These biomarkers are then used to train machine learning models, improving their interpretability and utility in clinical decision-making. Table 4 shows the biomarkers in order of greatest to least influence of this presented example . This step can also be presented as a percentage ranking format, which allows for more accurate emphasis on the weight of each feature. The Figures 1, 2 and 3 are represented in this format.

The PDF method effectively identifies subtle differences between patient groups, enabling the development of clinically relevant and interpretable machine learning models.

4. Datasets

To evaluate the effectiveness of the proposed **Pairwise Difference Filter (PDF)** method, we conducted experiments using three distinct datasets: the **MUSIC Dataset**, the

Feature	Importance
Age	0.179427
Weight (kg)	0.144872
Height (cm)	0.130086
BMI (kg/m^2)	0.132916
NYHA Class	0.149452
Diastolic Blood Pressure (mmHg)	0.198344

Table 4. Most Influence Features for This Example.

COVID-19 Dataset, and the **Wine (Toy Dataset)**. Each dataset represents a unique domain and presents specific challenges, allowing us to assess the robustness and generalizability of our method across different applications.

- The **MUSIC Dataset** focuses on cardiovascular signals and clinical data is used to validate approaches for predicting sudden cardiac death(SCD) and pump failure death (PFD) in patients with chronic heart failure (CHF) [Martin-Yebra et al. 2025].
- The **COVID-19 Dataset** contains also ECG and clinical data from patients with varying severity levels of COVID-19, enabling the evaluation of our method in a pandemic-related healthcare context [Pordeus 2023].
- The **Wine Toy Dataset** is a well-known benchmark dataset used for classification tasks, providing a controlled environment to test the method's ability to identify discriminative features in simpler, non-medical data [Dua and Graff 2019].

In the following subsections, we provide a detailed description of each dataset, highlighting their characteristics, relevance, and how they were used to validate the PDF method.

4.1. MUSIC

The **MUSIC (MUerte Subita en Insuficiencia Cardiaca)** dataset, available on PhysioNet, is a comprehensive resource for studying sudden cardiac death (SCD) and cardiac mortality in patients with chronic heart failure (CHF). The dataset includes clinical and electrocardiographic (ECG) data from 992 ambulatory CHF patients enrolled across eight Spanish hospitals between 2003 and 2004. Each patient underwent extensive evaluations, including 3-lead resting ECG, 24-hour Holter ECG (2-lead in 4% of patients and 3-lead in 96% of patients), echocardiography, chest X-ray, and blood laboratory tests. The dataset also provides long-term follow-up data, with a median follow-up period of 44 months, to assess outcomes such as sudden cardiac death (SCD) and pump failure death (PFD). The MUSIC dataset is widely used for developing predictive models and risk stratification tools in cardiovascular research, making it a valuable resource for understanding and preventing cardiac mortality in CHF patients [Martin-Yebra et al. 2025].

For our experiments, we used clinical and demographic information, radiographic, echocardiographic, laboratory variables and medications summarized in the CSV file subject-info.csv, focusing in the classification between classes "Sudden Cardiac Death" vs "Pump-Failure Death". For MUSIC dataset, we chose to select features until the sum of the percentage in the ranking reached 80%. With this, ten features were selected, 67% of the total (10 out of 15).

4.2. Covid-19

This dataset was obtained in quantitative cross-sectional descriptive study of a quantitative nature carried out at Hospital Universitário Walter Cantídio (HUWC) and Hospital Estadual Leonardo da Vinci (HELV), Ceará, Brazil, from May 2021 to January 2022, corresponding to the first, second and third wave of COVID-19 in Brazil. Participants were approached for convenience in isolation wards intended for the treatment of COVID-19. Individuals aged the same or over 18 years of age with a confirmed clinical diagnosis of COVID-19 by the RT-PCR test, and excluded individuals using invasive mechanical ventilation or non-invasive, using beta-blocker medication, oral or inhaled beta agonist and vasoactive drug, or with a history of syncope, pre-syncope or known arrhythmias. The study followed the ethical precepts of the Declaration of Helsinki [Association 2001]. The evaluations were only started after the patient had understood the protocol completely and written informed consent was obtained. The study was approved by the Ethics Committee for Research with Human Beings (CAAE - HELV: 47229221.9.3001.5684; CAAE - HUWC: 47229221.9.0000.5045)[Pordeus 2023].

The dataset consists of hospitalized individuals with clinical symptoms and a confirmed diagnosis of COVID-19. Patients were classified according to the severity of their condition, as shown in Table 5. The dataset includes 17 patients with low or moderate severity and 33 patients with severe symptoms. The ECG signals were collected following established protocols to minimize circadian influences and ensure signal stability [Pordeus 2023]. To demonstrate the usability and versatility of the PDF method, we used as dataset only the HRV metrics extracted using Neurokit [Makowski et al. 2021] and applied the method to them. We selected the features in the range of $[max, max + std]$ over the feature ranking, getting 36 out of 75 features, 48%.

Low or **Moderate**	Mild clinical symptoms and no signs of pneumonia on examination image, Presence of fever and respiratory symptoms with evidence radiology of pneumonia
Severe	Respiratory distress ($>$ 30 breaths / min) Oxygen saturation $<$ 93% at rest Partial pressure of arterial oxygen (PaO2)/ fraction of inspired oxygen (FiO2) $<$300 mmHg Cases with chest images that showed $>$50% evident lesion progression within 24-48 hours

Table 5. COVID-19 severity classification criteria [Yan 2020].

4.3. Wine (Toy Dataset)

The **Wine Toy Dataset** is a widely used benchmark dataset in machine learning, often used for classification tasks. It originates from the UCI Machine Learning Repository [Dua and Graff 2019] and consists of 178 samples of wine, categorized into three classes (Class 0, Class 1, and Class 2). Each sample is described by 13 continuous features derived from chemical analyzes, such as alcohol content, malic acid, and color intensity. The dataset is particularly useful for evaluating classification algorithms due to its small size, well-defined structure, and clear separability between classes. Its simplicity and

interpretability make it an ideal choice for testing preprocessing methods, feature selection techniques, and model interpretability in controlled experiments. For this dataset we selected 50% of the features.

5. Results

Images 1, 2 and 3 present the Ranking of Feature Importance for each dataset tested in this work. With it we can choose how much data to use to train the model. For this dataset we choose to select 50% of the features, getting 6 out of 13.

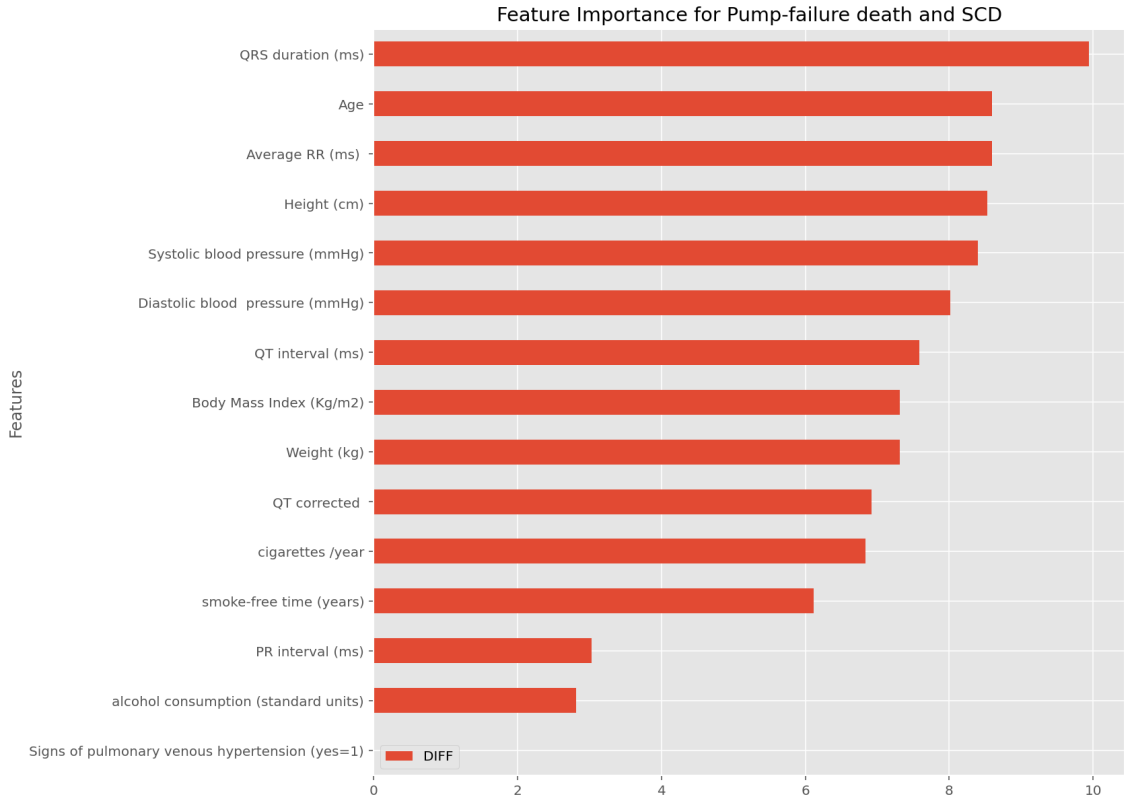

Figure 1. Feature Rank for MUSIC Dataset

We evaluated the performance of PDF on a COVID-19 severity classification dataset using several machine learning models, MUSIC classes Cardiac Sudden Death vs Pump-Failure Death, and Class 0 vs Class 1 for Wine Dataset. The results are summarized in Table 6. PDF achieved competitive performance while providing interpretable feature rankings that align with clinical knowledge.

In general, we got a slight performance loss, but this is expected when working with less training data. We have obtained gains in some cases, probably due to the elimination of outliers caused by the application of PDF and, as part of future work, we will check what could have caused the performance gain. In fact, performance gain was never the original goal in the development of this method.

6. Conclusions

Pairwise Difference Filter (PDF) provides a novel and effective approach to preprocessing, with a focus on enhancing medical interpretability. By identifying clinically mean-

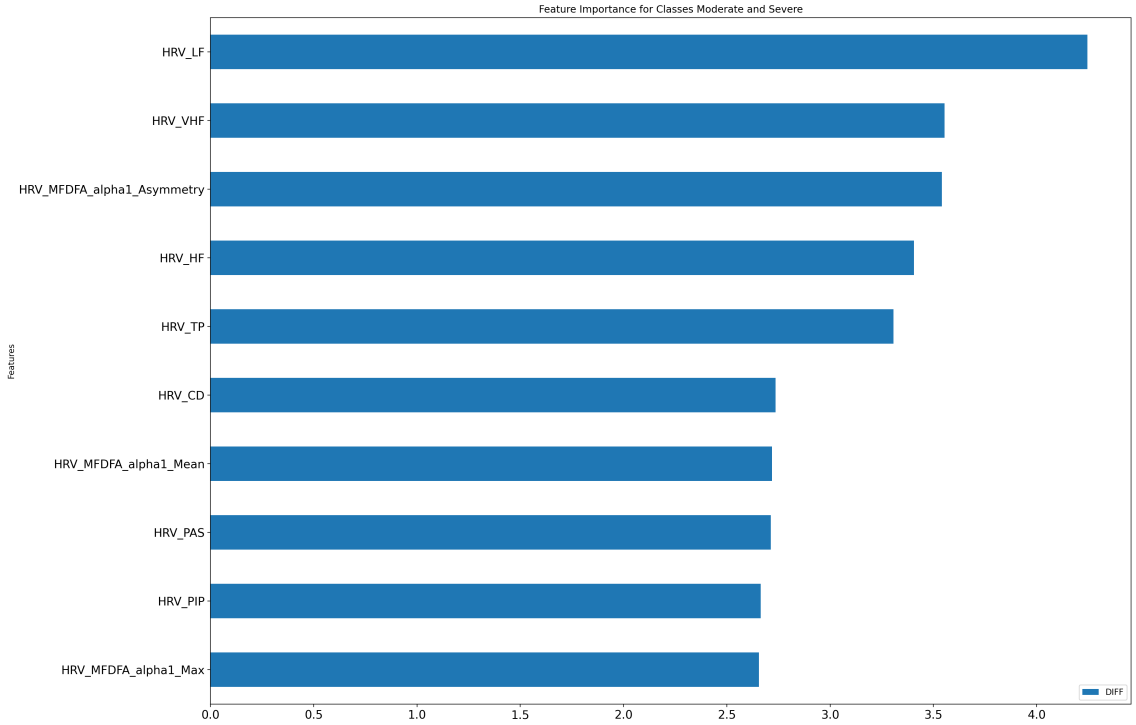

Figure 2. Feature Rank (First 10) for COVID-19 Dataset

Dataset	Model	PDF Acc.	PDF Recall	PDF Prec.	PDF F1	Comp. Acc.	Comp. Recall	Comp. Prec.	Comp. F1
MUSIC	RF	68	50	70	58	68	55	72	59
(SCD vs	GB	68	61	80	59	68	61	75	63
Pump-Fail.)	LR	66	55	80	59	63	55	100	57
COVID-19	RF	80	86	86	86	90	100	87	93
(Moderate	GB	80	86	100	86	90	100	87	93
vs Severe)	LR	80	100	100	83	80	100	78	87
Wine	RF	96	100	93	96	100	100	100	100
(Class 0 vs	GB	96	100	93	96	96	100	96	100
(Class 1)	LR	100	100	100	100	100	100	100	100

Table 6. Comparison of Metrics Using PDF and the Complete Dataset.

ingful biomarkers, PDF supports medical decision-making and improves the transparency of predictive models in healthcare, excelling in scenarios where interpretability is critical [Ribeiro et al. 2016].

The results indicates that PDF is a promising preprocessing method that can maintain the quality of the results, and before that, bring a preview information of what is happening with the data, showing different importance degrees for the features, bringing interpretability to nearly stages of analysis, without any machine learning method. Its performance is robust across different datasets and models, though its effectiveness may vary depending on the specific characteristics of the dataset and the model used. These findings suggest that PDF is a valuable tool for applications in healthcare and other do-

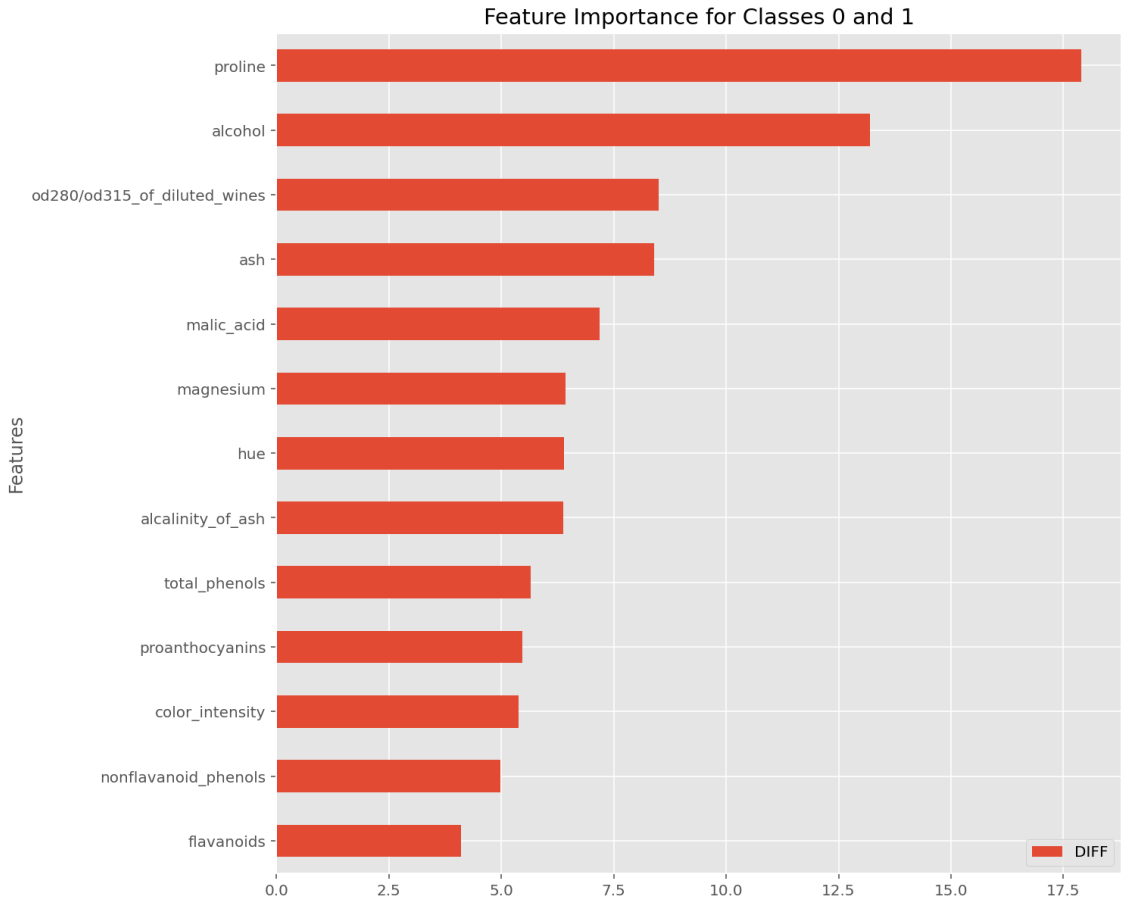

Figure 3. Feature Rank for Wine Toy Dataset

mains where both interpretability and performance are critical. The trade-offs between feature reduction and model performance, as well as the interpretability gains provided by PDF, would be beneficial to fully understand its potential.

Future work will focus on extending the method in ways that can use PDF to help predict clinical conditions. For example, MUSIC Class 0 are patients who have not died yet, but who have cardiac conditions that could lead to sudden death or pump failure. Having a good response to this prediction could help in the correct treatment of patients' conditions, improving their life expectancy.

Acknowledgments

This study was sponsored in part by FUNCAP BMD-0008-00739.01.02/24, CNPQ 420576/2023-1 and Petrobras S.A.

References

Association, W. M. (2001). World medical association declaration of helsinki: Ethical principles for medical research involving human subjects. *Bulletin of the World Health Organization*, 79(4):373–374.

Caldas, W. (2024). *IVS: INTERPRETATIVE VARIABLE SELECTION VIA PERFECT BIPARTITE MATCHING*. PhD thesis, Federal Unversity of Ceará.

Dhurandhar, A., Chen, P.-Y., Luss, R., Tu, C.-C., Ting, P., Shanmugam, K., and Das, P. (2018). Explanations based on the missing: Towards contrastive explanations with pertinent negatives.

Dua, D. and Graff, C. (2019). UCI machine learning repository.

Hakkoum, H., Abnane, I., and Idri, A. (2022). Interpretability in the medical field: A systematic mapping and review study. *Applied Soft Computing*, 117:108391.

Jović, A., Brkić, K., and Bogunović, N. (2015). A review of feature selection methods with applications. In *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 1200–1205.

Kira, K. and Rendell, L. A. (1992). The feature selection problem: Traditional methods and a new algorithm. *Proceedings of the Ninth International Workshop on Machine Learning*, pages 129–134.

Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324.

Kononenko, I. (1994). Estimating attributes: Analysis and extensions of relief. *European Conference on Machine Learning*, pages 171–182.

Lisboa, P., Saralajew, S., Vellido, A., Fernández-Domenech, R., and Villmann, T. (2023). The coming of age of interpretable and explainable machine learning models. *Neurocomputing*, 535:25–39.

Makowski, D., Pham, T., Lau, Z. J., Brammer, J. C., Lespinasse, F., Pham, H., Schölzel, C., and Chen, S. H. A. (2021). NeuroKit2: A python toolbox for neurophysiological signal processing. *Behavior Research Methods*, 53(4):1689–1696.

Martin-Yebra, A., Martínez, J. P., and Laguna, P. (2025). Music (sudden cardiac death in chronic heart failure). *PhysioNet*.

Pordeus, D. e. a. (2023). Training strategies for covid-19 severity classification. In Rojas, I., Valenzuela, O., Rojas Ruiz, F., Herrera, L., and Ortuño, F., editors, *Bioinformatics and Biomedical Engineering*, volume 13919 of *Lecture Notes in Computer Science*. Springer, Cham.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier.

Van Looveren, A. and Klaise, J. (2021). Interpretable counterfactual explanations guided by prototypes. In Oliver, N., Pérez-Cruz, F., Kramer, S., Read, J., and Lozano, J. A., editors, *Machine Learning and Knowledge Discovery in Databases. Research Track*, pages 650–665, Cham. Springer International Publishing.

Yan, X. e. a. (2020). Clinical characteristics and prognosis of 218 patients with covid-19: a retrospective study based on clinical classification. *Frontiers in medicine*, 7:485.