Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue on Recent Advances in Bayesian Statistics
and Machine Learning
Volume 24, No. 2, 2026 (New Series), pp 247–267
http://www.ssca.org.in/journal

A New Technique for AI Explainability using Feature Association Map

Sayantani Ghosh[1], Amit Kumar Das[2] and Amlan Chakrabarti[1,2,3]

[1]*DBS Bank*
[2]*Institute of Engineering & Management*
[3]*University of Calcutta*

Received: 04 June 2026; Revised: 15 August 2026; Accepted: 09 September 2026

Abstract

Lack of transparency in AI systems poses challenges in critical real-life applications. It is important to be able to explain the decisions of an AI system to ensure trust on the system. Explainable AI (XAI) algorithms play a vital role in achieving this objective. In this paper, we propose a new algorithm for explaining AI systems, FAMeX (Feature Association Map based eXplainability). The proposed Feature Association Map (FAM) algorithm combines graph-based feature redundancy with a classifier-dependent relevance measure computed from prediction probability changes after feature permutation. Consequently, FAMeX provides a classifier-specific post-hoc explanation instead of relying solely on dataset statistics. The proposed FAMeX algorithm combines feature redundancy represented through Feature Association Maps with classifier-specific feature relevance estimated through prediction probability changes under feature permutation. Experiments conducted with eight benchmark datasets compare FAMeX with Permutation Feature Importance (PFI) and SHapley Additive exPlanations (SHAP). The results show that FAMeX achieves higher predictive information retention in the feature-selection evaluation conducted in this study. This demonstrates that FAMeX is a promising framework for generating classifier-specific global feature importance explanations.

Key words: Feature association; Feature Importance; Explainable AI; Feature similarity; Feature relevance.

1. Introduction

For the last few decades, artificial intelligence (AI) based systems have been adopted in multiple facets of daily life. Whether we look at frequent applications like weather forecasting, prediction of stock prices, or we look at more advanced applications like early detection of diseases from medical images, analyse sentiment of texts, *etc.*, the role of AI systems is quite predominant. Major applications of AI systems, as we can see, are in the prediction of

Corresponding Author: Sayantani Ghosh
Email: sayantani31g@gmail.com

events, for which certain complex learning algorithms are employed. These complex learning algorithms can give superior, almost close to accurate, predictions.

The success of AI systems has also brought forward some deep-lying challenges. When we look at sophisticated AI algorithms, that albeit includes the complex learning algorithms, it is sometimes very difficult to understand how the algorithms are making decisions. It is sometimes impossible to understand whether the decision made by the AI system is based on scientific facts or spurious coincidences. This is fine when we work with non-critical systems like one which recommends books, movies or music to users. But when we go for critical implementations of AI like detection of diseases, forecasting weather, or predicting stock prices, the consequences might be fatal. So, for critical AI based decision systems, the opacity of the black-box AI system needs to be broken and an explanation of the decisions taken by the system should be evident. To foster trust within people, it is essential to conserve reliability, accountability, *etc.* to maintain transparency between computational systems and human beings Arrieta *et al.* (2020).

To make an AI system transparent and to explain why a particular algorithm takes a specific decision, the concept of explainability of AI systems has evolved. In the last few years, a lot of research has been done by the AI research community which is focused on establishing transparency of AI systems. Explainable AI (or popularly termed as XAI) algorithms [Chen *et al.* (2018); Doran *et al.* (2017); Tagaris and Stafylopatis (2020)] are intended to come up with certain explanations or understanding about how a particular AI system is taking decisions. XAI algorithms have seen a new apex in recent years [Doshi-Velez and Kim (2017); Dosilovic *et al.* (2018)]because conventional forecasting of AI decision-making systems need to be complimented with human-understandable rationale. This has been further necessitated by the fact that the more predictive a system is, the less is its interpretability (Das and Rad, 2020).

Broadly, XAI algorithms are based on few key philosophies. The first one is based on perceptible interpretation. These algorithms try to explain the decisions by AI systems using human perception-based intuitions. The other approach is to explain the AI systems by doing mathematical modelling of the system. A popular approach for explainability is based on feature importance. In this approach the XAI algorithm tries to represent feature importance in the form of ranking of features. There are a few algorithms proposed by the researchers to come up with the feature importance in context of decision-making of an AI system [Giudici and Raffinetti (2020); Merrick and Taly (2020)]. But they have not done a holistic analysis of the critical aspects of the features - relevance and redundancy. A graph-based formulation of the feature set, termed as Feature Association Map (FAM), proposed in our earlier works [Das *et al.* (2017); Dua and Graff (2019)] has been able to successfully model the aspects of features namely feature relevance and redundancy. We envisage that the same formulation of FAM can also be used to interpret feature importance in context of understanding a prediction done by an AI system in a more holistic way. The proposed FAMeX estimates relevance directly from the trained classifier through prediction perturbation while simultaneously accounting for redundancy.

The novelty of FAMeX is the integration of redundancy acquired by the Feature Association Map and classifier specific relevance determined by changes in prediction prob-

abilities after feature permutation. This implies that FAMeX is able to generate global post-hoc explanations and penalise redundant features simultaneously. Hence, it has the ability to measure the importance of contribution of the feature in the decision making process. FAM, a visual representation of the feature set, has been used while deriving the feature importance.

The remaining part of the paper has been organized in the following sections:

- Section 2 presents some of the existing methodologies of explainable AI which are relevant to our contribution.
- Section 3 highlights the underlying conceptual framework used in the proposed work.
- Sections 4 and 5 present the proposed methodology and FAMeX algorithm in detail.
- Extensive experimental results have been reported in Section 6. This section also contains the comparative study and validation of FAMeX.
- The work has been summarized in the form of conclusion in Section 7.

2. Related works

XAI techniques are adopted by users to establish trust on AI solutions. LIME Ribeiro *et al.* (2016) uses local models to explain its prediction, Deep LIFT Chen *et al.* (2018) explains the output with the help of backpropagation through the neurons of a network, Partial Dependence Plot **?** explains how the model's prediction partially depends on values of the input variables of interest.

Permutation Feature Importance (PFI) Huang *et al.* (2016) calculates the increase or decrease of the model's prediction error after randomly shuffling the values of a feature while retaining the feature in the dataset. This gives a measure of the feature importance. Among all the model agnostic local explanation techniques the most popular unified approach is SHAP typified for SHapley Additive exPlanations Merrick and Taly (2020) Redelmeier *et al.* (2020) Bowen and Ungar (2020) Parsa *et al.* (2019). It estimates the importance of each feature based on its marginal contribution. SHAP uses the concept of game theory that calculates the respective contribution of the players in a group (coalition). To interpret the learning models it estimates the contribution of each feature to a model's decision. In a work Veropoulos *et al.* (1999), SHAP is used to interpret and analyze the features obtained by the occurrence of accidents in Chicago Metropolitan Expressways.

Recent research has further advanced post-hoc explainability through improved evaluation methodologies, robustness analysis, and the study of feature attribution under correlated inputs. Several studies have investigated the faithfulness and reliability of feature attribution methods Asazuma *et al.* (2023). Furthermore, the limitations of permutation-based feature importance in the presence of correlated features have been analysed, and correlation-aware permutation strategies have been proposed to improve attribution robustness Kaneko (2022). Unlike these approaches, the proposed FAMeX explicitly incorporates feature redundancy through a graph-based Feature Association Map while estimating feature relevance from the prediction behaviour of the trained classifier, thereby producing

classifier-specific global feature importance explanations.

Minimum Redundancy Maximum Relevance (mRMR) Peng *et al.* (2005), ReliefF Kononenko (1994), and conditional mutual information-based approaches Fleuret (2004) are the classical filter-based feature selection methods. Their objectives and formulations differ from FAMeX because they do not require a trained predictive model and select features primarily from statistical relationships. In this paper, feature redundancy is modelled through feature associations, and feature relevance is derived from the prediction behaviour of a trained classifier through permutation-based prediction changes. Therefore, mRMR Peng *et al.* (2005), ReliefF Kononenko (1994), and conditional mutual information-based methods Fleuret (2004) are not considered as direct baselines in the present post-hoc, classifier-dependent evaluation. However, comparison with these filter-based approaches remains an important direction for future work.

3. Conceptual framework

3.1. Feature similarity

In a set of features, the degree of similarity between the features represents one aspect of feature association Chen *et al.* (2018) Bang *et al.* (2019) Das *et al.* (2016). Similarity expresses the redundancy between the set of features drawing a clear inference from the measured values. To measure the similarity between the features, Pearson's correlation coefficient is extensively used. For a pair of random variables X and Y, correlation coefficient r is defined by Equation 1.

$$r = \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}} \tag{1}$$

where

$$cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) \tag{2}$$

and

$$var(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{3}$$

$$var(Y) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \tag{4}$$

Correlation coefficients range between -1 and +1, where a maximum of linear correlation is indicated by +1, minimum correlation by -1 and no correlation is indicated by 0. A range of values between 0.68 and 1.00 is considered as high. Moreover, values larger than 0.9 indicate very high correlation Mukaka (2012).

3.2. Feature relevance

FAMeX computes feature relevance directly from the trained classifier. The objective is to quantify how much the prediction of the trained model depends on an individual feature.

For every feature, its values are randomly permuted while keeping all other features unchanged. The prediction probabilities of the trained classifier before and after permutation are compared. A larger prediction difference indicates that the model relies heavily on that feature during decision making, whereas a smaller difference indicates lower relevance.

The relevance of feature f_i is calculated as

$$\Delta_i = \frac{1}{N} \sum_{k=1}^{N} \left| P_M(y|x_k) - P_M(y|x_k^{perm(i)}) \right| \tag{5}$$

where

- $P_M(y|x_k)$ denotes the prediction probability of the trained classifier for sample x_k,
- $x_k^{perm(i)}$ denotes the same sample after randomly permuting feature f_i, and
- N is the total number of samples.

The relevance score is then obtained by normalizing the prediction difference as

$$Relevance\ Score_i = \frac{\Delta_i}{\frac{1}{m}\sum_{j=1}^{m} \Delta_j} \tag{6}$$

where m is the total number of features.

3.3. Graph clustering

A graph $G = \{E, V, C\}$ is considered to have a set of edges $E = \{e_1, e_2, \ldots\}$, vertices $V = \{v_1, v_2, \ldots\}$ and colors $C = \{c_1, c_2, \ldots\}$ where each element of the objects is colored in $\{c_1, c_2, \ldots\}$ and the edges e_i are associated with unordered pairs of vertices $\{v_1, v_2, \ldots\}$ Pavlopoulos *et al.* (2004) Nahar and Ara (2018) Parthiban *et al.* (2011). Vertices which are densely connected or form a sub-graph in a graph can be discovered from graph clustering.

4. Proposed methodology

The opaque AI systems furnish limited understanding about the input-output relationship. This lack of interpretability is attempted to be addressed with our proposed algorithm. The proposed algorithm Feature Association Map based eXplainability (or FAMeX) comes up with a feature importance based explanation of the system output. The foundation of the FAMeX algorithm is the graph-theoretic concept to model feature association, which is termed as Feature Association Map (FAM). FAM, as shown in Fig. 1, is a graph that has features modelled as vertices and the association between the features represented as edges.

There are two facets of this association - one is relevance and the other is redundancy. Relevance signifies the degree to which a trained classifier depends on a particular feature while making predictions. Instead of estimating relevance solely from the statistical properties of the dataset, FAMeX evaluates the influence of each feature by measuring the change

in the prediction probability of the trained model after randomly permuting that feature. Features producing larger prediction changes are considered more relevant for explaining the model's decision. So more information a particular feature contributes, the more significant or relevant it is.

Again, multiple features can contribute information in terms of deciding the class value. But certain features might be contributing similar information. Those features can be regarded as potentially redundant. Potentially redundant features can be considered to be less important in context of the decision taken by an AI system.

Unlike conventional filter-based feature selection methods, relevance score depends explicitly on the trained classifier. Therefore different classifiers trained on the same dataset may produce different relevance scores and consequently different feature rankings. This makes FAMeX a classifier-specific post-hoc explanation framework.

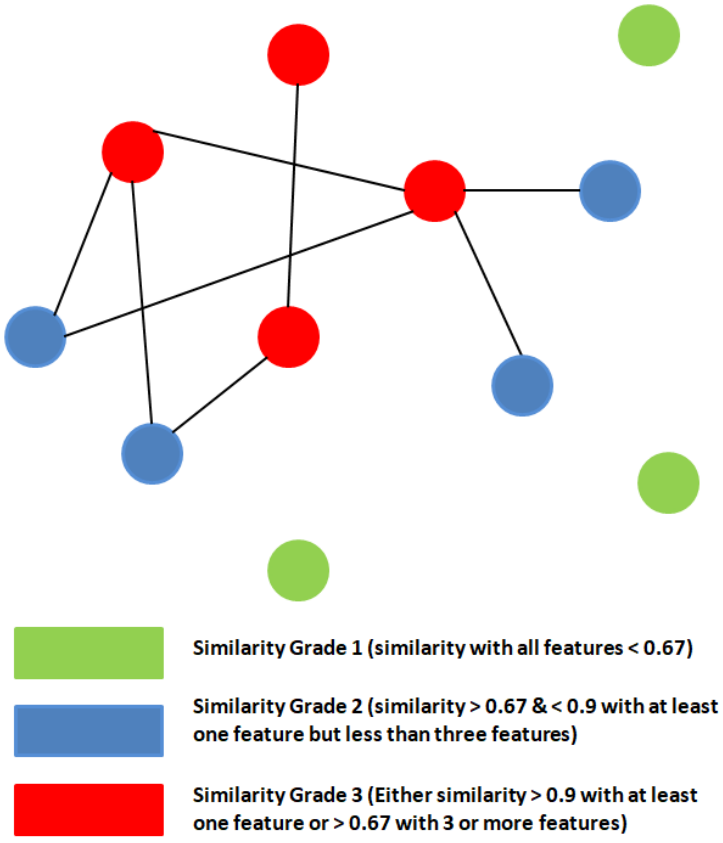

Figure 1: FAMeX Feature Graph

In the proposed FAMeX algorithm, the relevance and redundancy level of the features are evaluated. As has been presented in Fig. 1, the overall *feature importance score* is calculated based on two measures - (a) the *similarity score* and (b) the *relevance score*. Based on certain criteria highlighted below, the features obtained from FAM Das *et al.* (2017) are graded in three levels - Grade 1, Grade 2, and Grade 3. The deciding factors underlying the grades are:

- Grade 1 is assigned to features which have low redundancy, *i.e.*, correlation value less than 0.67 with all other features. Feature similarity grade assigned for these features is 1.
- Grade 2 is assigned to features which have moderate redundancy, *i.e.*, correlation value more than 0.67 but less than 0.9 with at least one feature but less than three features. Feature similarity grade assigned for these features is 2.
- Grade 3 is assigned to features which have high redundancy, *i.e.*, correlation value more than 0.9 with at least one feature or more than 0.67 with three or more features. Feature similarity grade assigned for these features is 3.

Although correlations of 0.68 and 0.89 both belong to Grade 2, they represent features that exhibit moderate redundancy and therefore receive the same redundancy penalty. The proposed grading scheme aims to distinguish broad levels of redundancy rather than determining the exact magnitude of correlation. The exact contribution of each feature is subsequently differentiated through the classifier-dependent relevance score, which is computed continuously from prediction probability changes. Consequently, the final feature importance score remains continuous even though the redundancy component is discretised.

Based on this similarity grade, the similarity score is calculated for each feature using Equation 9.

$$Similarity\ Score_i = \frac{(Feature\ Similarity\ Grade_i)^2}{Average\ Feature\ Similarity\ Grade} \tag{7}$$

The objective of the proposed grading is not to estimate the exact magnitude of redundancy but to distinguish among low, moderate and high redundancy regions. Since FAM is a feature association graph, the redundancy information is represented using discrete vertex grades rather than continuous edge weights. This enables a simple graph representation while preserving the overall redundancy structure of the feature set.

The threshold of 0.67 corresponds approximately to the transition from moderate to strong positive correlation according to widely used statistical interpretation guidelines, whereas correlations above 0.90 are generally regarded as indicating very strong redundancy. Consequently, these thresholds provide a practical separation between weakly associated, moderately redundant and highly redundant features.

Next, the prediction dependency of every feature is estimated by measuring the average change in prediction probability of the trained classifier after randomly permuting that feature. The obtained prediction differences are normalized to compute the relevance score. Relevance score for each feature is calculated below in Equation 8.

$$Relevance\ Score_i = \frac{\Delta_i}{\frac{1}{m}\sum_{j=1}^{m}\Delta_j} \tag{8}$$

To calculate feature importance score for each feature, similarity score and the relevance score are integrated in such a way that the score of the feature is higher if the relevance

score is higher and the score of the feature is lower if the similarity score is higher. Thus, similarity score has an inverse relation whereas relevance score has a direct relation with the feature importance score.

$$Feature\ Importance\ Score_i = \frac{Relevance\ Score_i}{Similarity\ Score_i} \tag{9}$$

Thus, the explanation is derived from both the structural characteristics of the dataset and the prediction behaviour of the trained classifier. The similarity score captures redundancy among correlated features, whereas the relevance score captures the dependency of the trained model on individual features. Their combination enables FAMeX to generate classifier-specific post-hoc explanations.

The top features are the ones that have the highest feature importance scores. These features are expected to have greater explanatory importance for the prediction behaviour of the trained classifier.

The similarity grade is an ordinal measure representing three levels of redundancy, namely low, moderate and high similarity. Since highly redundant features contribute substantially less additional information than weakly correlated features, the penalty should increase non-linearly rather than linearly. Therefore, the similarity score is defined using a quadratic function of the similarity grade. The quadratic formulation amplifies the penalty assigned to highly redundant features while preserving the ordering of the three similarity levels. Consequently, Grade 3 features receive a substantially larger penalty than Grade 2 or Grade 1 features, encouraging the selection of features that contribute complementary rather than duplicated information.

5. Algorithm

Algorithm 1 FAMeX Algorithm

Require: Training dataset D, trained classifier M, feature set $F = \{f_1, f_2, \ldots, f_m\}$
Ensure: Feature importance scores $I = \{I_1, I_2, \ldots, I_m\}$

```
M_corr ← |correlation(D_F)|
for each element of M_corr do
    if colnum = rownum then
        M_corr[colnum, rownum] ← 0
    end if
    if M_corr[colnum, rownum] ≥ 0.67 then
        M_corr[colnum, rownum] ← 1
    else
        M_corr[colnum, rownum] ← 0
    end if
end for
```

Algorithm 1 FAMeX Algorithm (continued)

Construct graph $G_{\text{FAM}} \leftarrow \text{adjacencyGraph}(M_{\text{corr}})$
for each feature $f_i \in F$ **do**
  Determine the number of features having correlation ≥ 0.67 with f_i
  Determine the number of features having correlation ≥ 0.90 with f_i
  if number of correlated features ≥ 0.90 is zero **and** number correlated at ≥ 0.67 is zero **then**
    $nodegrade_i \leftarrow 1$
  else if number correlated at ≥ 0.90 is at least one **or** number correlated at ≥ 0.67 is at least three **then**
    $nodegrade_i \leftarrow 3$
  else
    $nodegrade_i \leftarrow 2$
  end if
end for
for each feature $f_i \in F$ **do**
  Randomly permute values of f_i
  Compute prediction probabilities before and after permutation
  $\Delta_i \leftarrow \frac{1}{N} \sum_{k=1}^{N} \left| P_M(y|x_k) - P_M(y|x_k^{perm(i)}) \right|$
end for
Compute $\overline{\Delta} \leftarrow \frac{1}{m} \sum_{j=1}^{m} \Delta_j$
for each feature $f_i \in F$ **do**
  $relscore_i \leftarrow \dfrac{\Delta_i}{\overline{\Delta}}$
  $simscore_i \leftarrow \dfrac{(nodegrade_i)^2}{\frac{1}{m} \sum_{j=1}^{m} (nodegrade_j)^2}$
  $I_i \leftarrow \dfrac{relscore_i}{simscore_i}$
end for
return I

6. Experiments and results

6.1. Selection of baseline explainability methods

Several explainability methods have recently been proposed, including Local Interpretable Model-agnostic Explanations (LIME) Ribeiro *et al.* (2016), Integrated Gradients (IG) Sundararajan *et al.* (2017), and Anchors Ribeiro *et al.* (2018). However, these approaches primarily generate *local explanations* for individual predictions or require differentiable neural network architectures. In contrast, the objective of the proposed FAMeX framework is to compute a *global feature importance ranking* that can be applied to arbitrary machine learning classifiers.

Therefore, the experimental evaluation focuses on two widely accepted model-agnostic explainability techniques, namely Permutation Feature Importance (PFI) Kaneko (2022) and SHAP Lundberg and Lee (2017). PFI is inherently a global explanation method, whereas

SHAP can generate both local (instance-level) and global explanations. Since FAMeX aims at providing global feature importance scores, we compare the scores obtained by aggregating SHAP values over the dataset. Thus, all three methods are compared under the same global explanation setting.

FAMeX generates a global post-hoc explanation by estimating the overall importance of each feature across the evaluation dataset. It is intended to explain the general prediction behaviour of a trained classifier rather than individual predictions. Consequently, the comparison with SHAP in this work considers the aggregated global feature importance scores produced by SHAP rather than its instance-level explanations.

6.2. Experimental setup

Eight benchmark datasets, presented in Table 1, have been considered for the experiments. All the datasets are available publicly in the University of California (UCI) Machine Learning repository Akyol and Gültepe (2017). These datasets are obtained from various domains to check the variability in the outcomes of the FAMeX and other competing algorithms.

Table 1: Details of the datasets used in experiments

Datasets	No. of Instances	No. of Features
Wisconsin	683	10
ILPD	580	11
Pageblocks	5474	11
Pima	769	9
Apndcts	107	8
WineQuality	1600	12
WBDC	570	31
Vehicle	847	19

We have used Google Colaboratory platform for programming and Python 3.7 as the programming language.

For every classifier, FAMeX recomputes the relevance score using the prediction probabilities produced by that classifier. Therefore each classifier generates its own feature ranking. This makes the comparison with SHAP and PFI fair since all three methods explain the same trained model. This comparative study can provide a better understanding of how the proposed FAMeX algorithm performs in comparison to the existing benchmark algorithms with the same datasets.

The relevance score in the algorithm is computed independently for every trained classifier using the prediction probability change after feature permutation. Consequently, FAMeX produces classifier-specific feature importance rankings, allowing a fair comparison with other post-hoc explainability techniques such as SHAP and PFI.

6.3. Evaluation metrics

To evaluate the performance of the proposed FAMeX algorithm compared to the competing SHAP and PFI algorithms, feature importance based ranking has been considered. To examine the predictive information retained by the rankings, the most important features identified by each algorithm have been used to derive a subset of the dataset. Classification accuracy has then been calculated using standard classifiers.

This evaluation should be interpreted as an indirect assessment of whether highly ranked features retain predictive information. It does not constitute a direct measure of explanation quality, faithfulness, or human interpretability.

We have also taken a view by selecting the least important features. For a given algorithm, it is expected that the data subset generated by the most important features should retain more predictive information than the data subset generated by the least important features.

Robust classifiers, namely Support Vector Machine (SVM) Alty *et al.* (2004)Pavlopoulos *et al.* (2004), Random Forest (RF) Subasi *et al.* (2017), Naive Bayes (NB) Parthiban *et al.* (2011), Ramadhan *et al.* (2020), and Decision Tree (DT) Pavlopoulos *et al.* (2004), Nahar and Ara (2018), are used to evaluate the predictive performance of the selected feature subsets. All the experiments have been repeated for 100 iterations.

For the experiments presented in this paper, the top 30% of the FAMeX features and bottom 30% of the FAMeX features are extracted and classified. The same procedure is applied to PFI and SHAP rankings. The resulting classification accuracies are used only to assess predictive information retention within the selected subsets.

In addition to the average classification accuracy, statistical significance of the performance differences between FAMeX, SHAP, and PFI was evaluated using the Wilcoxon signed-rank test across the benchmark datasets. A significance level of $\alpha = 0.05$ was adopted.

In this paper, classification accuracy is considered as an indirect validation of the feature importance rankings, rather than a direct measure of the quality of explanation provided by the XAI approach. The top-ranked features having higher predictive performance indicate that they retain substantial predictive information. This evaluation should not be interpreted as evidence of explanation faithfulness or consistency.

6.4. Results and analysis

The FAM formation for the WineQuality dataset has been presented in Fig. 2. As can be observed, 5 out of 11 features are represented by green vertices, indicating low similarity / redundancy. Only one feature, fixed.acidity, has similarity grade 3.

Abbreviations: M = Mean, W = Worst (Largest (worst) value observed for that feature), SE = Standard Error, Perim = Perimeter, ConcPts = Concave Points, Smooth = Smoothness, Adh. = Adhesion, Chrom. = Chromatin, Nuc. = Nucleoli, Sz. = Size.

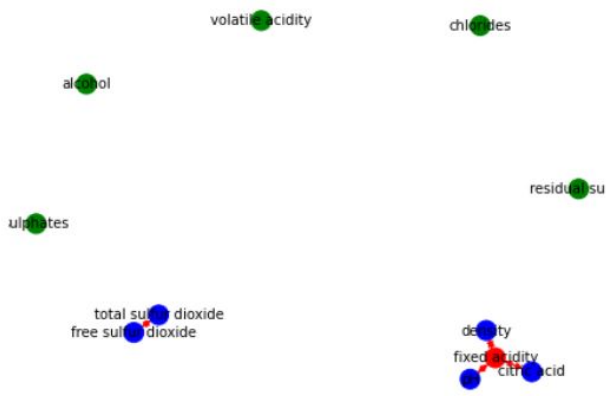

Figure 2: Association Map for WineQuality dataset

Table 2: Top-ranked features identified by FAMeX (classifier-specific), PFI and SHAP for the Wisconsin dataset.

Rank	FAMeX			PFI	SHAP
	RF	DT	SVM		
1	Area_W	ConcPts_W	Area_W	Clump Thick.	Bare Nuclei
2	Radius_W	Perim_W	Area_M	Mitosis	Bland Chrom.
3	ConcPts_W	Area_M	Perim_W	Bland Chrom.	Clump Thick.
4	ConcPts_M	Texture_M	Area_{SE}	Bare Nuclei	Epithelial Sz.
5	Perim_W	Smooth_W	Texture_W	Marginal Adh.	Cell Shape
6	Area_{SE}	Texture_W	Texture_M	Cell Shape	Cell Size
7	Concavity_M	Area_{SE}	Perim_M	Normal Nuc.	Mitosis
8	Area_M	Concavity_W	Radius_W	Epithelial Sz.	Marginal Adh.
9	Concavity_W	Area_W	Perim_{SE}	Cell Size	Normal Nuc.
10	Radius_M	ConcPts_M	Radius_M	—	—

Table 2 compares the Top-10 important features identified by PFI, SHAP and the proposed FAMeX for the Wisconsin Breast Cancer dataset. While PFI and SHAP generate a single feature ranking for the dataset, FAMeX produces classifier-specific rankings because the relevance scores are computed from the prediction behaviour of the trained classifier.

It can be observed that the rankings generated by Random Forest, Decision Tree and SVM are not identical, indicating that each classifier relies on a slightly different subset of features during prediction. Nevertheless, several important features such as *area_worst*, *perimeter_worst*, *radius_worst* and *concave points_worst* consistently appear among the highest-ranked features across multiple classifiers, demonstrating that FAMeX identifies stable predictive attributes while preserving classifier-specific decision characteristics.

Unlike PFI and SHAP, which provide a single global ranking, FAMeX adapts its explanations to the underlying learning algorithm. This property is further examined through the Spearman rank correlation analysis presented in the subsequent section.

6.4.1. Consistency of feature rankings across classifiers

Since the proposed FAMeX estimates feature relevance from the prediction behaviour of the trained classifier, different classifiers are expected to generate different feature importance rankings. To quantify the degree of agreement among these rankings, Spearman's Rank Correlation Coefficient (ρ) was computed between the complete feature rankings generated by Random Forest, Naive Bayes and Support Vector Machine (SVM).

Spearman correlation measures the similarity between two ranked lists. A value close to 1 indicates that the classifiers produce highly similar rankings, whereas values close to 0 indicate weak agreement.

The correlation matrix obtained is presented in Table 3. All pairwise correlations are statistically significant ($p < 0.001$), indicating that while the classifiers assign different relative importance to some features, they largely agree on the overall ordering of important features.

Random Forest and SVM ($\rho = 0.784$) has the highest similarity, while for Naive Bayes it is comparatively lower with both Random Forest ($\rho = 0.689$) and SVM ($\rho = 0.662$).

Hence, FAMeX produces explanations that remain largely stable and consistent across different classifiers while reflecting each classifier's individual decision-making characteristics.

Table 3: Spearman rank correlation between FAMeX feature rankings generated by different classifiers

	Random Forest	Naive Bayes	SVM
Random Forest	1.000	0.689	0.784
Naive Bayes	0.689	1.000	0.662
SVM	0.784	0.662	1.000

Table 4: Statistical significance of Spearman rank correlations

Classifier Pair	Spearman ρ	p-value
Random Forest vs Naive Bayes	0.689	2.6×10^{-5}
Random Forest vs SVM	0.784	$< 10^{-6}$
Naive Bayes vs SVM	0.662	6.8×10^{-5}

The statistically significant positive correlations in Table 4 demonstrate substantial agreement among FAMeX feature rankings obtained using different classifiers. Although the relevance score is computed independently for each trained model, the important features identified by different classifiers remain substantially consistent. At the same time, the observed differences in ranking indicate that FAMeX retains classifier-specific information rather than producing a fixed dataset-dependent ranking.

6.4.2. Sensitivity analysis of correlation thresholds

Three different correlation threshold settings, $(0.60, 0.85)$, $(0.67, 0.90)$, and $(0.75, 0.95)$, were considered for sensitivity analysis. The consistency of feature rankings was quantified using Spearman's rank correlation coefficient (ρ) relative to the default threshold setting $(0.67, 0.90)$, and the corresponding Random Forest classification accuracy was recorded.

Table 5: Sensitivity analysis of the proposed correlation thresholds.

Threshold Setting	Spearman's ρ	RF Accuracy (%)
0.60 / 0.85	0.991	96.49
0.67 / 0.90 (Default)	1.000	96.49
0.75 / 0.95	0.867	96.49

As shown in Table 5, the feature rankings remain highly consistent across all threshold settings, with Spearman's rank correlation ranging from 0.867 to 1.000 relative to the default configuration. Furthermore, the Random Forest classification accuracy remains unchanged at 96.49% for all three settings.

These findings demonstrate that the proposed framework is relatively robust to the tested variations in the selected correlation thresholds, indicating that the exact threshold values have limited influence on the resulting feature rankings and predictive performance for this experiment.

6.4.3. Classification performance using SVM, random forest, naive Bayes and decision tree

Table 6: Accuracy (using SVM) with top 30% and tottom 30% important features

	Top 30%			Bottom 30%		
Dataset	FAMeX	PFI	SHAP	FAMeX	PFI	SHAP
Wisconsin	96.1±1%	95.22±1%	93.93±1%	94.22±1%	89.58±1%	89.24±1%
ILPD	87.61±1%	71.19±1%	71.06±1%	85.21±1%	71.25±1%	71.62±1%
Pageblocks	94.86±2%	89.85±2%	89.24±2%	89.91±2%	89.67±2%	90.10±2%
Pima	82.74±1%	66.12±1%	75.45±1%	81.2±1%	64.88±1%	64.64±1%
Apndcts	88.9±1%	80.5±1%	83.09±1%	86.34±1%	85.68±1%	81.22±1%
WineQuality	71.16±1%	49.98±1%	49.01±1%	68.10±1%	55.41±1%	46.11±1%
WBDC	86.66±2%	85.22±2%	85.35±2%	62.98±2%	92.85±2%	92.51±2%
Vehicle	59.76±2%	40.69±2%	41.07±2%	52.7±2%	37.20±2%	36.77±2%
Average	**83.47±1%**	**72.35±1%**	**73.53±1%**	**77.58±1%**	**72.07±1%**	**71.53±1%**

Let's first start with the results of the SVM classifier presented in Table 6. As can be observed, FAMeX achieves higher classification accuracy than both competing methods for most of the datasets. With the top 30% features, FAMeX achieves an average accuracy of 83.47%, compared to 72.35% by PFI and 73.53% by SHAP.

For most of the datasets, the classification accuracy generated by the top 30% features identified by FAMeX is higher than that generated by the bottom 30% features. For Pima, WineQuality and WBDC, the difference is particularly pronounced.

Table 7: Accuracy (using RF) with top 30% and bottom 30% important features

	Top 30%			Bottom 30%		
Dataset	FAMeX	PFI	SHAP	FAMeX	PFI	SHAP
Wisconsin	95.67±1%	93.77±1%	94.29±1%	92.67±1%	89.66±1%	89.04±1%
ILPD	72.54±1%	69.98±1%	68.13±1%	71.76±1%	63.87±1%	68.53±1%
Pageblocks	94.57±2%	90.71±2%	90.08±2%	92.26±2%	90.34±2%	91.22±2%
Pima	67.79±1%	59.44±1%	67.37±1%	66.54±1%	58.50±1%	61.42±1%
Apndcts	83.45±1%	78.45±1%	78.54±1%	83.61±1%	83.31±1%	79.0±1%
WineQuality	67.38±1%	52.94±1%	52.17±1%	66.57±1%	57.35±1%	49.86±1%
WBDC	89.02±2%	82.39±2%	83.42±2%	80.44±2%	90.35±2%	90.21±2%
Vehicle	72.76±2%	44.1±2%	44.52±2%	78.19±2%	38.25±2%	53.50±2%
Average	**80.52±1%**	**71.47±1%**	**71.82±1%**	**79.005±1%**	**71.45±1%**	**72.85±1%**

For the Random Forest classifier, as shown in Table 7, FAMeX achieves higher top-30% classification accuracy than PFI and SHAP for most of the datasets. Overall, the top-30% features identified by FAMeX achieve an average accuracy of 80.52%, compared to 71.47% by PFI and 71.82% by SHAP.

The results indicate that the top-ranked FAMeX features generally retain substantial predictive information. However, some datasets, such as WBDC and Vehicle, show cases in which the bottom-ranked subset can also retain considerable predictive information. This illustrates that feature redundancy and predictive information can coexist in the feature space.

Table 8: Accuracy (using NB) with top 30% and bottom 30% important features

	Top 30%			Bottom 30%		
Dataset	FAMeX	PFI	SHAP	FAMeX	PFI	SHAP
Wisconsin	95.14±1%	95.04±1%	93.32±1%	89.75±1%	85.62±1%	88.98±1%
ILPD	72.56±1%	50.48±1%	69.07±1%	70.59±1%	68.36±1%	49.98±1%
Pageblocks	91.08±2%	78.21±2%	87.50±2%	90.22±2%	78.35±2%	90.45±2%
Pima	76.95±1%	67.98±1%	74.92±1%	76.62±1%	65.55±1%	66.46±1%
Apndcts	88.48±1%	66.59±1%	83.77±1%	84.37±1%	84.99±1%	65.36±1%
WineQuality	61.11±1%	49.31±1%	48.3±1%	61.03±1%	55.83±1%	40.39±1%
WBDC	92.5±2%	82.13±2%	82.39±2%	86.42±2%	90.11±2%	90.78±2%
Vehicle	62.1±2%	31.09±2%	31.13±2%	59.61±2%	53.5±2%	38.27±2%
Average	**79.99±1%**	**65.10±1%**	**71.3±1%**	**77.32±1%**	**72.78±1%**	**66.33±1%**

As presented in Table 8, for the Naive Bayes classifier, FAMeX achieves higher top-30% classification accuracy than both PFI and SHAP for most of the datasets. Overall, the top-30% features identified by FAMeX achieve an average accuracy of approximately 80%, compared to 65.10% by PFI and 71.30% by SHAP.

The difference between top-30% and bottom-30% FAMeX features is particularly notable for ILPD, WineQuality and WBDC.

Table 9: Accuracy (using DT) with top 30% and bottom 30% important features

	Top 30%			Bottom 30%		
Dataset	FAMeX	PFI	SHAP	FAMeX	PFI	SHAP
Wisconsin	95.8±1%	93.48±1%	93.86±1%	90.61±1%	89.16±1%	88.82±1%
ILPD	76.12±1%	66.36±1%	65.68±1%	74.11±1%	62.56±1%	66.90±1%
Pageblocks	94.56±2%	90.72±2%	90.12±2%	90.78±2%	90.70±2%	91.37±2%
Pima	70.12±1%	57.48±1%	65.72±1%	69.19±1%	59.06±1%	61.50±1%
Apndcts	83.56±1%	75.86±1%	70.95±1%	81.10±1%	77.09±1%	75.59±1%
WineQuality	66.04±1%	53.17±1%	52.72±1%	64.12±1%	56.75±1%	49.80±1%
WBDC	92.22±2%	79.87±2%	80.97±2%	89.24±2%	88.38±2%	89.07±2%
Vehicle	59.38±2%	43.41±2%	43.29±2%	58.23±2%	53.91±2%	53.67±2%
Average	**79.725±1%**	**75.11±1%**	**70.41±1%**	**77.17±1%**	**72.20±1%**	**70.84±1%**

As shown in Table 9, the results for the Decision Tree classifier show a generally favourable pattern for FAMeX. For the top 30% features, FAMeX achieves an average accuracy of 79.725%, compared to 75.11% by PFI and 70.41% by SHAP.

However, the results also show that FAMeX does not outperform both baselines for every individual dataset and feature subset. For example, for the Pageblocks dataset using the bottom 30% features, SHAP achieves 91.37%, compared with 90.78% for FAMeX. Such cases highlight that the evaluation measures predictive information retention rather than directly establishing explanation quality.

Table 10: Average Accuracy (all classifiers) with top 30% and bottom 30% important features

	Top 30%			Bottom 30%		
Classifier	FAMeX	PFI	SHAP	FAMeX	PFI	SHAP
SVM	83.47±1%	72.35±1%	73.53±1%	77.58±1%	72.07±1%	71.53±1%
RF	80.52±1%	71.47±1%	71.82±1%	79.005±1%	71.45±1%	72.85±1%
NB	79.99±1%	65.10±1%	71.3±1%	77.32±1%	72.78±1%	66.33±1%
DT	79.725±1%	75.11±1%	70.41±1%	77.17±1%	72.20±1%	70.84±1%

The summary of results of all experiments has been presented in Table 10. As can be observed, FAMeX achieves the highest average top-30% accuracy among the three methods for all four classifiers. The average accuracy obtained using the top 30% FAMeX features is 83.47% for SVM, 80.52% for RF, 79.99% for NB, and 79.725% for DT.

The difference between top-30% and bottom-30% FAMeX subsets is positive for all four classifiers, although the magnitude of the difference varies across classifiers and datasets. In contrast, PFI and SHAP do not show the same consistently positive separation across all classifier–dataset combinations.

These results indicate that FAMeX rankings retain substantial predictive information

in the selected top-ranked features. However, classification accuracy should be interpreted only as an indirect evaluation of the feature ranking and not as a direct measure of explanation faithfulness or trustworthiness.

Table 11: Wilcoxon signed-rank test comparing Top 30% classification accuracies across benchmark datasets

Classifier	Comparison	Statistic	p-value
Decision Tree	FAMeX vs PFI	0.0	0.0078
	FAMeX vs SHAP	0.0	0.0078
Naive Bayes	FAMeX vs PFI	0.0	0.0078
	FAMeX vs SHAP	0.0	0.0078
Random Forest	FAMeX vs PFI	0.0	0.0078
	FAMeX vs SHAP	0.0	0.0078
SVM	FAMeX vs PFI	0.0	0.0078
	FAMeX vs SHAP	0.0	0.0078

The Wilcoxon signed-rank test results presented in Table 11 indicate statistically significant paired differences between FAMeX and both PFI and SHAP across the eight benchmark datasets for each classifier ($p = 0.0078 < 0.05$). These results support the observation that FAMeX produces higher predictive information retention in the top-ranked feature subsets under the evaluation protocol adopted in this study.

It should be emphasized that this statistical test evaluates differences in the selected predictive-performance metric and therefore does not directly establish the faithfulness, reliability, or human interpretability of the explanations.

6.5. GUI-based tool

A Graphical User Interface (GUI) based tool has been developed which integrates the implementation of the FAMeX algorithm. The tool allows the user to browse a dataset. The tool generates the similarity score, relevance score and feature importance score of the features of the selected dataset along with the classification accuracy for the four standard classifiers - SVM, RF, NB and DT.

The percentage of top or bottom features can be taken as user input. This tool is platform-independent. It assists in understanding the important features of a particular dataset in the context of prediction. This prototypic tool has been uploaded in the GitHub repository `https://github.com/Sayantanighosh17github/FAMeX-Tool`, so that it can be used by the AI community.

7. Conclusion

In this paper, a new algorithm, FAMeX, has been proposed for generating classifier-specific global feature importance explanations. The algorithm integrates graph-based fea-

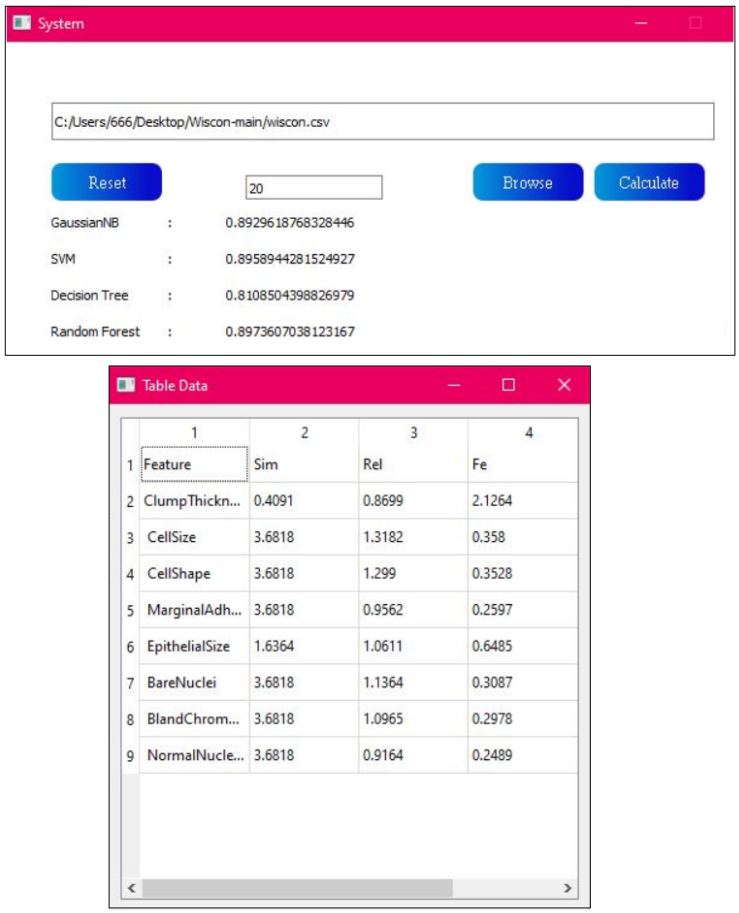

Figure 3: FAMeX Feature Detector Tool

ture redundancy with classifier-dependent relevance estimated through changes in prediction probabilities after feature permutation.

The experimental evaluation across eight benchmark datasets shows that the top-ranked FAMeX feature subsets generally retain more predictive information than the corresponding bottom-ranked subsets. FAMeX also achieves higher average top-30% classification accuracy than PFI and SHAP across the four evaluation classifiers considered in this study.

The results demonstrate the usefulness of combining feature redundancy and model-dependent relevance within a unified feature importance framework. However, the classification-based evaluation should be interpreted as an indirect assessment of predictive information retention rather than as direct evidence of explanation faithfulness or trustworthiness.

Although a linear penalty could also be employed, preliminary experiments indicated that the quadratic formulation provides better discrimination between moderately and highly redundant features, leading to more stable feature rankings. A comprehensive comparison with alternative penalty functions, additional explainability metrics, and larger datasets remains an interesting direction for future work.

8. Future direction

As a possible future work, uncertainty can be incorporated into the feature importance estimation. Relevance score depends on the prediction probability after feature permutation, so its variability can be studied through repeated experiments across different datasets. To quantify this uncertainty, a Bayesian approach can help estimate the relevance and feature importance scores, which may provide more reliable feature rankings Johnsen *et al.* (2023).

Future work can also investigate explanation faithfulness, stability under distributional perturbations, alternative redundancy penalties, conditional permutation strategies for correlated features, and evaluation on larger and more complex datasets. Comparison with additional model-agnostic and model-specific XAI techniques would further establish the generality of the proposed framework.

Conflict of interest

The authors declare that they have no competing interests or conflicts of interest in connection with this work. No specific funding was received for this study or any related activities.

Data availability

The data used in this study are from open-source repositories and publicly available sources, as detailed in Section 6.1 of this manuscript. All information and datasets referenced are accessible to the public and do not require additional permissions for access.

References

Akyol, K. and Gültepe, Y. (2017). A Study on liver disease diagnosis based on assessing the importance of attributes. *International Journal of Intelligent Systems and Applications*, **9**, 1–9.

Alty, S. R., Millasseaut, S. C., Chowienczyk, P. J., and Jakobsson, A. (2004). Cardiovascular disease prediction using support vector machines. *IEEE*, 0-7803-8294-3/04/$20.00.

Arrieta, A., Díaz-Rodríguez, N., Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *ArXiv*, 1910.10045.

Asazuma, Y., Hanawa, K., and Inui, K. (2023). Empirical analysis of methods for evaluating faithfulness of explanations by feature attribution. *Transactions of the Japanese Society for Artificial Intelligence*, **38**, C–N22_1–C–N22_9.

Bang, S., Xie, P., Wu, W., and Xing, E. (2019). Explaining a black-box using deep variational information bottleneck approach. *ArXiv*, *abs/1902.06918*.

Bowen, D. and Ungar, L.(2020). Generalized SHAP: Generating multiple types of explanations in machine learning. *ArXiv*, *abs/2006.07155*.

Breiman, L. (2001). Random forests. *Machine Learning*, **45**, 5–32.

Chen, J., Song, L., Wainwright, M., and Jordan, M. I. (2018). Learning to explain: An information-theoretic perspective on model interpretation. *ArXiv*, *abs/1802.07814*.

Das, A., Goswami, S., Chakrabarti, A., and Chakraborty, B. (2017). A new hybrid feature selection approach using feature association map for supervised and unsupervised classification. *Expert Systems with Applications*, **88**, 81–94.

Das, A., Goswami, S., Chakraborty, B., and Chakrabarti, A. (2016). A graph-theoretic approach for visualization of data set feature association. *ACSS*.

Dua, D. and Graff, C. (2019). *UCI Machine Learning Repository*. University of California, Irvine, School of Information and Computer Sciences. `https://archive.ics.uci.edu`

Das, A. and Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (XAI): A Survey. *ArXiv*, *abs/2006.11371*.

Doran, D., Schulz, S., and Besold, T.R. (2017). What does explainable AI really mean? A new conceptualization of perspectives. *ArXiv*, *abs/1710.00794*.

Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv: Machine Learning*.

Dosilovic, F. K., Brčič, M., and Hlupic, N. (2018). Explainable artificial intelligence: A survey. *41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 0210-0215.

Fleuret, F. (2004). Fast binary feature selection with conditional mutual information. *Journal of Machine Learning Research*, **5**, 1531–1555.

Johnsen, P. V., Strømke, I., Langaas, M., DeWan, A. T., and Riemer-Sørensen, S. (2023). Inferring feature importance with uncertainties with application to large genotype data. *PLoS Computational Biology*, **19**, e1010963.

Giudici, P., and Raffinetti, E. (2020). Shapley-Lorenz decompositions in explainable artificial intelligence. *Econometrics: Econometric Model Construction*, SSRN 3546773.

Goldstein, A., Kapelner, A., Bleich, J., and Pitkin, E. (2014). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *arXiv*, *1309.6392*.

Huang, N., Lu, G., and Xu, D. (2016). A permutation importance-based feature selection method for short-term electricity load forecasting using random forest. *Energies*, **9**, 767.

Kaneko, H. (2022). Cross-validated permutation feature importance considering correlation between features. *Analytical Science Advances*, **3**, 278–287.

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

Lundberg, S. M. and Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. In: *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 30.

Merrick, L. and Taly, A. (2020). *The Explanation Game: Explaining Machine Learning Models using Shapley Values. CD-MAKE*.

Mukaka, M. M. (2012). A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal*, **24**, 69.

Nahar, N. and Ara, F. (2018). Liver disease prediction by using different decision tree techniques. *International Journal of Data Mining and Knowledge Management Process*, **8**, 01-09.

Parsa, A., Movahedi, A., Taghipour, H., Derrible, S., and Mohammadian, A. (2019). Toward safer highways: Application of XGBoost and SHAP for real-time accident detection and feature analysis. *Accident Analysis and Prevention*, **136**, 105405.

Parthiban, G., Srivatsa, S., and Rajesh, A. (2011). Diagnosis of heart disease for diabetic patients using Naive Bayes method. *International Journal of Computer Applications*, **24**, 7–11.

Pavlopoulos, S., Stasis, A., and Loukis, E. (2004). A decision tree–based method for the differential diagnosis of aortic stenosis from mitral regurgitation using heart sounds. *BioMedical Engineering OnLine*, **3**, 21.

Peng, H., Long, F., and Ding, C. (2005). Feature Selection Based on Mutual Information Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27**, 1226–1238.

Ramadhan, B., Purwanto, Y., and Ruriawan, M. F. (2020). Forensic malware identification using Naive Bayes method. *International Conference on Information Technology Systems and Innovation (ICITSI)*, 1-7.

Redelmeier, A., Jullum, M., and Aas, K. (2020). Explaining predictive models with mixed features using Shapley values and conditional inference trees. *ArXiv*, *abs/2007.01027*.

Ribeiro, M.T., Singh, S., and Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

Subasi, A., Alickovic, E., and Kevric, J. (2017). Diagnosis of chronic kidney disease by using random forest. In *CMBEBIH 2017: Proceedings of the International Conference on Medical and Biological Engineering*, 589–594.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In: *Proceedings of the 34th International Conference on Machine Learning*, 3319–3328.

Tagaris, T. and Stafylopatis, A. (2020). Hide-and-Seek: A template for explainable AI. *ArXiv*, *abs/2005.00130*.

Veropoulos, K., Cristianini, N., and Campbell, C. (1999). The application of support vector machines to medical decision support: A case study. *ACAI99*.

Zhou, Y., Booth, S., Ribeiro, M. T., and Shah, J. (2022). Do feature attribution methods correctly attribute features? In *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*, 9623–9633.