

Received 9 July 2024, accepted 29 August 2024, date of publication 3 September 2024, date of current version 12 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3454084

RESEARCH ARTICLE

ARIA, HaRIA, and GeRIA: Novel Metrics for Pre-Model Interpretability

MAREK PAWLICKI 

Faculty of Telecommunications, Computer Science and Electrical Engineering, Bydgoszcz University of Science and Technology, 85-796 Bydgoszcz, Poland
ITTI Sp. z o.o., 61-612 Poznań, Poland

e-mail: marek.pawlicki@pbs.edu.pl

This work was supported by the Horizon Europe AI4CYBER Project through European Union's Horizon Europe Research and Innovation Programme under Grant 101070450.


ABSTRACT This work proposes three novel Pre-Model Interpretability metrics: HaRIA, ARIA, and GeRIA. They aim to assess the potential utilization of features in machine learning models prior to the training phase, by quantifying the Relative Information Availability. These metrics integrate Mutual Information and ANOVA F-values, scaled using Maximum Absolute Scaling. This allows to evaluate the potential of a feature being used in the learning process efficiently and effectively without the computational expense of model training. The metrics are designed to provide a holistic view of feature relevance by capturing both the non-linear dependencies and variance effects among features. Validation of these metrics across multiple datasets demonstrates their capability to approximate the importance assigned by more complex models, as evidenced by their strong correlation with traditional feature importance measures and SHAP values obtained post-model training. The consistency observed in various datasets underscores the potential of RIA metrics to facilitate early-stage model development decisions, offering a cost-effective tool for feature evaluation in scenarios where computational resources are limited or rapid prototyping is necessary. However, some discrepancies, especially with complex models like ANNs, indicate areas for future research and refinement. The introduction of these metrics marks a significant step toward enhancing the efficiency and transparency of AI development by enabling a better understanding of data characteristics and potential model behavior before actual model deployment.

INDEX TERMS Data characteristics, feature relevance, feature utilization, machine learning, model behavior, pre-model interpretability.

I. INTRODUCTION

In the beginning, there is data. Having good data allows one to perform different learning tasks. These are broadly categorized as supervised learning, unsupervised learning, and reinforcement learning [1]. The focus of this paper is on supervised learning in classification.

The main idea in supervised Machine Learning (ML) is to use data to make models that can determine something (Y) based on some information (X). To perform its task, an ML algorithm summarizes X to the form of T . The best summary T should give us as much information about Y as possible

The associate editor coordinating the review of this manuscript and approving it for publication was Bing Li .

while keeping only the important parts of X [2]. How well an ML algorithm can learn depends on two aspects: model complexity and training samples - both their amount and the quality of the information they carry. The ML methods need a number of samples to train the model to obtain good outputs.

The Vapnik-Chervonenkis (VC) dimension measures how complicated a model can get. It looks at how many different ways one can separate or classify the available data points. If the model can make a lot of different splits, it has a high VC dimension. In practice, when the data provides a multitude of features, the model becomes more complex. But more complexity does not always translate into better performance. A model should be 'just right', e.g. not too simple, and not too complex [3].

To achieve better performance, practitioners attempt to reduce complexity, by feature selection, dimensionality reduction, or by adjusting hyperparameters, like the number of neurons, layers, trees, etc., to find the right balance [3]. Additionally, regularization techniques are employed to prevent overfitting [4]. This helps ensure the model generalizes well to new, unseen data. Finally, effective model training employs cross-validation, where the data is split into multiple subsets to validate the model's performance [5]. By balancing complexity and training data effectively, ML models can achieve robust performance, offering reliable results. However, as ML and AI technologies started being employed in crucial areas of human endeavor, the users realized that the results were not the only things that mattered, and more issues emerged: the security of deployed AI models and the fact that many of them work as a black box, without providing any clue on how they reached the decisions [6]. An explainable/interpretable AI (xAI) system has an ability to demystify its processes, providing the why and the how of its decisions in a language understandable to its observers. While currently there is an ongoing debate on the terminology - drawing distinctions between closely related terms such as "explainability," "interpretability", "understandability", "comprehensibility", "intelligibility", and "transparency" [7], [8], in this paper the terms will be used interchangeably - so as not to take a side in the argument until the terms coagulate.

Explainability is crucial for several reasons [9], [10]:

- It helps identify and eliminate the "Clever Hans" models, where, as only the results are visible but not the process, the process is a hack. This transparency accelerates development and reduces the risk of faulty models.
- xAI ensures reliability by allowing regular verification of models and data, addressing issues like Concept Drift and Data Decay.
- It also builds trust in AI systems, especially in high-stakes decisions like medical diagnoses, where understanding the reasoning behind a decision is essential.
- xAI reduces bias and promotes fairness by revealing unwanted biases in datasets, aligning AI models with modern ethical standards.
- It provides deeper insights into the domain, uncovering unknown relationships within the data and facilitating new discoveries, thus proving invaluable to the scientific community.

These aspects of xAI are deemed important in the subject literature.

One cannot help but notice that all those points are just as important before any model is trained.

A. PROBLEM STATEMENT

All factors contributing to the significance of xAI retain their importance throughout the initial stages of model development, prior to the onset of any training.

It is crucial to identify features in the data that may lead the model to make decisions based on spurious correlations rather than genuine insights. This involves analyzing the relevance and contribution of each feature to potential outcomes.

Prior to training, verifying the integrity of the data and the assumptions underlying the modeling approach is essential. This includes checking for data quality, consistency, and distributions that match the expected patterns.

Before deploying ML models, it is important to rectify any biases in the dataset. This can involve statistical analyses to identify and address disproportionate representations or prejudiced patterns in the data, ensuring that the model training process starts from an equitable foundation.

Analyzing the relationships between features and their impact on the target variables can reveal complex interactions and dependencies, which are invaluable for domain understanding.

And finally, establishing mechanisms for transparency starts with a clear understanding of the characteristics of the data, and how they are expected to influence model behavior.

It is logical that addressing these aspects before model training not only streamlines the development process but also enhances the effectiveness, fairness, and reliability of AI systems.

Thus, a conundrum emerges: similar requirements which provide the domain pressure to develop methods to open the black box of AI, are present even before any model is trained. This constitutes pre-model interpretability (PMI) - methods of understanding the characteristics of data, and the potential behavior of ML models without the computational cost of training them. Traditional post-hoc approaches, like SHAP or feature importance scores, require a trained model. They look at the model, which is fitted to the data; however, without understanding what information was available in the data in the first place, there is no way to compare these metrics against a ground truth.

B. PRE-MODEL INTERPRETABILITY

An understanding of how a potential model could behave before undertaking training is particularly beneficial in the early stages of project feasibility studies, during dataset exploration, or when training is costly and computationally intensive, and when selecting features for high-stakes applications where interpretability and understanding are as crucial as performance.

This would constitute another step in the ML pipeline outlined at the beginning of this section, PMI would be in the vicinity of exploratory data analysis (EDA) and feature selection (FS). There would be some overlap with both EDA and FS: both PMI and EDA focus on gaining insights into the data before proceeding with modeling. Both EDA and PMI would use visual tools to uncover relationships between variables and understand data structure, and descriptive statistics would play a crucial role in both EDA and PMI. However, while EDA is used to form hypotheses and guide subsequent steps of analysis, including data cleaning and

transformation, PMI specifically aims to understand how features may influence model behavior and decisions. This not only involves looking at data distributions but also assessing potential feature importance and interactions in the context of modeling, which helps in estimating the predictive power of each feature, and subsequent model evaluation.

EDA is more about using intuition and visual tools to understand the data. It is an open-ended process used to make sense of data characteristics. PMI is more structured in its approach and focuses on understanding how data features will interact with ML algorithms, to predict how features could affect model outcomes. The insights gained from EDA are generally used to modify the dataset (e.g., removing outliers, handling missing values) and choose appropriate transformations. In PMI, the insights are used to anticipate how well features will convey information in a model, and ensure that the models built are interpretable and aligned with business goals.

C. PRE-MODEL INTERPRETABILITY, EXPLORATORY DATA ANALYSIS AND FEATURE SELECTION

It is important to emphasize that the primary goal of PMI is to understand data characteristics and the potential behavior of models on this data, before any model is built or trained. It aims to provide insights into the dataset itself, helping to anticipate how different models might interpret or weigh features. As such, it is distinct from feature selection, which aims to reduce the number of input variables to use in the final model. FS focuses on improving model performance by removing irrelevant or redundant features that could lead to overfitting or unnecessarily complex models. FS directly impacts model efficiency and effectiveness, often driven by the need to optimize computational resources and improve model accuracy, robustness, and generalization. FS employs algorithms and techniques such as recursive feature elimination, feature importance scores, and selection criteria based on model performance. PMI focuses on the understanding and explanation of feature roles and relationships without necessarily aiming to optimize any particular model's performance - rather to be able to provide a ground truth, an initial understanding of what is in the data. To be able to compare the understanding of data with what the model is doing after training. It is important to distinguish PMI from FS and acknowledge PMI relationship to FS, especially the filter methods. In fact, the proposed metrics stem directly from two filter methods - Mutual Information (Information Gain) and ANOVA. However, the unique perspective of the application of PMI is in its capacity to provide a foundational understanding before any predictive modeling occurs. This early-stage insight can influence the choice of the model, guide feature engineering efforts, and shape the overall strategy for data analysis. PMI thus acts as a critical step in ensuring that the models developed are not only effective but also understandable and aligned with the specific objectives of the project.

Ultimately, PMI, EDA and FS all aim to enhance the quality of the ML models, but they do so from different stages and different perspectives of the model development process. The position of PMI in relation to EDA and FS is showcased in Fig. 1. PMI helps to set a strong foundation for all subsequent ML steps, ensuring that the modeling is done on a well-understood and appropriately structured dataset.

D. RESEARCH QUESTIONS

With this background, the following research questions are formulated:

- RQ1 Can potential feature utilization in an ML model be reliably assessed before any model training?
- RQ2 Can pre-model interpretability metrics provide a reliable estimation of feature importance that aligns with traditional post-model training evaluations?
- RQ3 Can the proposed pre-model interpretability metrics provide a reliable estimation of what features will be utilized by the trained model, that aligns with traditional post-hoc xAI evaluations?

E. MAJOR PAPER CONTRIBUTIONS

To answer these questions, this paper offers the following major contributions:

- 1) The proposition of three PMI metrics, which incorporate two filter methods: Mutual Information (MI) and ANOVA F-values, scaled with Maximum Absolute Scaling and combined, to estimate feature importance effectively before model training. These metrics aim to blend the strengths of different statistical measures to provide an estimated evaluation of feature importance free from the peculiarities of model-specific metrics. Their usefulness in the PMI context is assessed and stacked against feature importance scores and SHAP values.
- 2) The proposed metrics were tested and validated across a variety of datasets, demonstrating their capability to align well with traditional model-based metrics, but without the cost of model training. This validation suggests that these metrics can provide reliable estimates similar to those derived from computationally expensive model training processes.
- 3) The paper provides a statistical analysis, showing how the proposed pre-model metrics correlate with traditional post-model feature importance scores and xAI metrics across multiple datasets.
- 4) The paper outlines potential future research to include synergistic relationships between features. This would address scenarios where the combined knowledge of two or more features significantly enhances the predictability of a target variable, beyond what could be inferred from each feature individually.
- 5) The proposed metrics offer a novel approach by assessing potential feature utilization before any model training occurs. By doing this, they provide a baseline

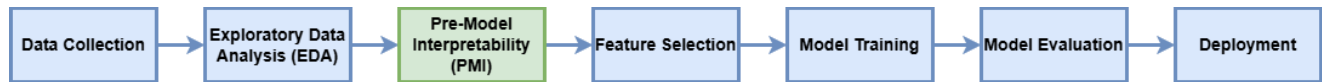


FIGURE 1. PMI in relation to the ML pipeline.

of what features are inherently important based on their information content and variance, independent of any specific model biases or constraints. This baseline could serve as a ground truth for comparing how post-training model explanations (like SHAP values), and feature importance scores align with what is inherently expected from the data. This objective baseline is crucial for xAI as it provides a non-biased point of reference to judge whether the explanations generated by xAI methods post-training truly reflect the underlying data dynamics or are artifacts of the modeling process.

- 6) By extension, an overview of feature importance free of model artifacts can help pinpoint if the modeling process goes according to the domain intuition and, by the same token, by divorcing the evaluation of data from the modeling process, help spot discrepancies in the data itself.

By providing insights into feature relevance before the expenditure of computational resources on model training, the proposed metrics allow for more efficient data analysis workflows. They might be particularly valuable in scenarios requiring rapid prototyping or where computational resources are constrained. Thus, the research addresses a significant gap in the machine learning workflow.

This research, via the proposed metrics, not only aims to conserve computational resources but also enhances the efficiency of the model development process by enabling better-informed decisions early in the development lifecycle. By providing this early insight, the metrics contribute to a more transparent, trustworthy, and equitable AI development process. This might be especially important in domains where understanding the basis of model decisions is critical. Thus, this research not only aims to provide tools for ML practitioners, but also to advance the discourse on xAI by advocating the divorcing of model explanations from the pre-model explanations, which analyze the information content of data the model will fit to.

F. PAPER STRUCTURE

The paper is structured as follows: Section II reviews the related work, Section III describes the materials and methods, detailing the statistical tools and algorithms used, including Mutual Information, ANOVA F-values, and ML models like Random Forests and Artificial Neural Networks. Section IV introduces the proposed metrics, explaining their components and computation. Section V outlines the experimental setup, including data preparation, model training, and the metrics used for evaluating feature importance. Results are presented in Section VI, showing the effectiveness of the proposed

metrics through various visualizations and statistical analyses. Finally, the conclusion in Section VII summarizes the findings and discusses future directions for research in enhancing pre-model interpretability metrics

II. RELATED WORKS

A. FEATURE SELECTION TECHNIQUES

In [11], a novel feature selection methodology aimed at enhancing model prediction accuracy in high-dimensional datasets is proposed. The method mixes the Relief filter algorithm with the multi-criteria decision-making method TOPSIS (Technique for Order Preference by Similarity to Ideal Solution) to form a new filter approach that models feature selection as a multi-criteria decision problem. Through the Relief method, a decision matrix is generated, which is then utilized by TOPSIS to rank features from the most to the least informative, facilitating the identification of the most significant features while avoiding overfitting.

The authors of [12] propose a novel hybrid optimization technique that combines the sine-cosine algorithm (SCA) with Harris hawks optimization (HHO) to tackle the challenges of feature selection in both low and high-dimensional datasets. This method, known as SCHHO, aims to enhance the exploration capabilities of HHO using the trigonometric operations of SCA, thereby improving global search efficiency and convergence speed without incurring additional computational costs. The approach is thoroughly evaluated through extensive experiments on the CEC'17 test suite and sixteen diverse datasets with over 15000 features.

In [13], the author provides a comprehensive overview of feature selection methods used in data mining and machine learning, particularly under the challenges and opportunities presented by big data. The paper reviews various types of feature selection algorithms, categorizing them primarily into similarity-based, information-theoretical-based, sparse-learning-based, and statistical-based methods. It highlights how feature selection facilitates simpler and more efficient models, enhances performance, and prepares cleaner data. The author also discusses the evolution of these methods to accommodate different data structures like streaming data, structured data, and heterogeneous data. Further, the survey introduces an open-source repository that includes many of these algorithms, promoting further research and application in the field.

The authors of [14] design a feature selection method for high-dimensional data, especially suited for classification tasks. The SA-EFS method combines the results of three different feature selection algorithms—chi-square test, maximum information coefficient, and XGBoost—through sort aggregation strategies like arithmetic and geometric mean

to determine the most important features. This ensemble approach helps overcome biases inherent in using a single feature selection method, leading to improved classification accuracy.

In [15], the authors research a dynamic feature selection method specifically designed for clustering high-dimensional data streams. This method, which utilizes a dynamic feature mask (DFM), addresses the challenges posed by feature drift, where the relevance of features can change as the data stream evolves. The paper emphasizes that traditional static feature selection methods are inadequate for data streams due to their inability to adapt to changing feature relevance over time. The proposed method is algorithm-independent and can be integrated with any existing density-based clustering algorithm, enhancing their performance by dynamically adjusting the feature set used for clustering.

The authors of [16] explore a feature selection method that combines feature grouping with Variable Neighborhood Search (VNS) techniques to address challenges in high-dimensional datasets. The approach involves using Markov blankets to group features and a metaheuristic search strategy to optimize feature selection. This method is tested across various datasets, including those from microarray and text mining, showing improved effectiveness over traditional methods by reducing the dimensionality while maintaining or enhancing model accuracy.

The survey contained in [17] provides a detailed review of various feature selection methods particularly applicable to big data environments. It categorizes these methods based on their nature, search strategies, and evaluation criteria, offering a structured taxonomy that highlights distinctions among similarity-based, information-theoretical, sparse-learning, and statistical approaches.

Another survey in [18] presents a detailed overview of feature selection methods employed in data mining and machine learning. It explores the importance of feature selection in handling high-dimensional data and improving the performance of mining algorithms by focusing on relevant features while discarding redundant ones. The study categorizes feature selection methods into three main types: Filter, Wrapper, and Hybrid methods, each with distinct mechanisms and applications.

The authors of [19] evaluate the effectiveness of various filter methods for feature selection across 16 high-dimensional datasets from diverse fields like bioinformatics. The study assesses 22 filter methods in terms of their runtime and predictive accuracy. Key findings include the absence of a universally superior group of filter methods, although some consistently perform well across multiple datasets.

B. INTERPRETABILITY AND EXPLAINABILITY IN MACHINE LEARNING

The author of [20] examines the philosophical and methodological dilemmas facing the field of interpretable machine learning (IML). Watson discusses how IML seeks to make

complex algorithms understandable to users, highlighting three main conceptual challenges: ambiguous fidelity to the target model versus the data generating process, lack of error rate control in IML methods, and the predominant focus on static explanations over dynamic processes. He argues that most IML tools fail to adequately address these challenges, which can lead to misleading or unintuitive explanations.

In [21], the fundamental principles for creating interpretable machine learning models are outlined. The authors argue that interpretability is essential for high-stakes decision-making and effective troubleshooting. The paper dispels common misconceptions that interpretability necessarily compromises model accuracy, presenting cases where interpretable models perform as well as or better than complex, less interpretable models. The ten challenges discussed range from optimizing sparse logical models like decision trees, to enhancing scoring systems, to incorporating constraints into generalized additive models, and to developing interpretable reinforcement learning techniques.

The authors of [22] evaluate various methods and dimensions of interpretability, emphasizing the importance of making these models understandable and trustworthy, especially when applied to critical domains like medicine and autonomous driving. It categorizes existing interpretability approaches into model transparency, functionality, and post-hoc explanations, discussing their strengths and limitations. Moreover, the paper identifies significant gaps in current methodologies, including the lack of a standard definition of interpretability, the variability of human expertise, the fact that there is a need for methods that integrate interpretability directly into the model training process, and that there is a lack of robust metrics to measure the effectiveness and accuracy of interpretations.

In [23], one can find an extensive review of the research on neural network interpretability. The survey proposes a novel taxonomy for interpretability organized along different dimensions, like type of engagement (passive vs. active interpretation approaches), type of explanation, and focus (ranging from local to global interpretability). One of the future directions identified by the paper is that research could focus on incorporating interpretability directly into the neural network training process, rather than treating it as an afterthought.

The authors of [24] provide a thorough investigation of Explainable Artificial Intelligence (xAI) within the context of cybersecurity, particularly focusing on deep learning and AI techniques. It is a comprehensive overview of xAI methods, emphasizing their significance and benefits in cybersecurity. The authors systematically map the potential future research directions through a methodical study, identifying and exploring the integration possibilities of xAI in cybersecurity. This investigation compiles various insights and concludes with the potential trajectories for future research and the integration of xAI into cybersecurity practices, making significant strides towards improving transparency, trust, and efficiency in AI-driven security systems.

The authors of [25] emphasize the need for machine learning algorithms to provide insights into their decision-making processes to ensure fairness, identify biases, and verify performance expectations. The paper explores various methods and challenges in developing effective explanations for AI systems, particularly in deep learning. It critiques the lack of standardization in explanations and suggests future research directions to improve transparency and trust in AI applications.

The authors of [26] are the first ones to point out that given that the foundation of any model is the data it is built upon, comprehending the dataset is crucial for both explainability and interpretability in the context of pre-model explainability. The paper presents techniques designed to provide insights into datasets for building more effective ML models. It reviews traditional methods such as univariate and multivariate exploratory data analysis (EDA) techniques.

III. MATERIALS AND METHODS

This section outlines the tools and methods employed to explore and validate the proposed metrics.

A. MUTUAL INFORMATION (MI)

is an information-theory quantity expressing how much the knowledge of one variable can reduce the uncertainty of another variable. MI of zero characterizes independent variables, and the higher the value, the more reduction in uncertainty one variable can provide about the other.

MI quantifies the highest possible amount of information to be extracted from one variable in order to decrease uncertainty about another [27]. This is important since, as it will be shown in the experiments, ML algorithms rarely utilize all the information available to them, as indicated by MI.

B. ANOVA F-VALUE

The F-statistic, or the F-value is a ratio of the mean squares of two variances, reflecting the various sources of variation in a set of data. In ANOVA F-test it is expressed as a ratio of between-group variability to the within-group variability. In the context of feature selection, ANOVA F-value is used to assess the significance of each feature in relation to the target variable. Features with high F-values contribute more to differentiating between the classes [28], [29]. The ANOVA F-value is one of the filter methods, a go-to method of feature selection.

C. RANDOM FOREST (RF)

is an ensemble, supervised ML algorithm which excels in both classification and regression tasks. The method constructs an array of decision trees, trains them on different subsets of the data, and then aggregates their outputs to estimate the final outcome. Each decision tree is trained on a random sample and a random selection of features. This technique - bootstrap aggregation (bagging) - inhibits overfitting and boosts the ability to generalize. On inference, the algorithm pushes each test sample through every decision

tree and uses the majority vote to determine the final outcome [30], [31]. The sci-kit learn implementation of RF provides insights into the significance of each feature, calculated through the average and standard deviation of the impurity reduction across each tree in the ensemble [32].

D. EXTRA TREES (ET)

classifier builds multiple unpruned decision trees using random selections of attributes and split points for each node. This extreme randomization reduces variance more effectively than traditional methods and ensures computational efficiency. The strength of randomization can be adjusted with a parameter, making the algorithm robust across different datasets [33]. In the scikit-learn implementation [32] of ET, the attribute *feature_importances* delivers the relative importance of features within the model, quantified by the normalized total reduction in Gini Impurity.

E. MAXIMUM ABSOLUTE SCALING

is a scaling technique based on the maximum absolute value observed in the set. This method ensures that the maximal absolute value of each feature is adjusted to 1.0, without shifting or centering the data. This preserves inherent sparsity. MaxAbs Scaling is a fundamental preprocessing step in ML, bringing numerical data to a common scale. For positive values, MaxAbs Scaler brings the data to the [0,1] range, which enhances the comparability across different variables [32].

F. ARTIFICIAL NEURAL NETWORKS

The astounding versatility of Artificial Neural Networks (ANNs) has been demonstrated across an abundance of fields. From their inception [34], through the innovation of Convolutional Neural Networks (CNNs) [35] and Recurrent Neural Networks (RNNs) [36], these models have become indispensable in applications like data mining, biometrics, natural language processing, shape recognition, face recognition, and intrusion detection, among many others. Today, neural networks are a backbone in many domains requiring rigorous data analysis.

The foundational idea behind ANNs is in their partial mimicry of biological neural networks [37]. This mimicry enables ANNs to recognize and extract meaningful patterns from data, a crucial capability when dealing with vast datasets.

ANNs refine their understanding of data by adjusting weights and biases through training. This involves iterating over data batches and tweaking parameters to minimize the error between the network's outputs and the actual outcomes, a process guided by a supervisory signal [38].

G. SHAPLEY ADDITIVE EXPLANATIONS (SHAP)

SHAP values are designed to attribute an importance value to each feature for a specific prediction, rooted in game theory's Shapley values, ensuring that feature attributions

are consistent and locally accurate. The framework proposed in [39] unifies several existing methods. The authors propose new approaches to calculating SHAP values, for better computational performance and better alignment with human intuition. The SHAP values reported in this paper are not local explanations but aggregated total absolute contributions across all the features estimated via SHAP over the entire training set.

IV. PROPOSED NEW METRIC: RELATIVE INFORMATION AVAILABILITY

The proposed “Relative Information Availability” (RIA) metrics are a set of three composite metrics that combine different aspects of features to provide an understanding of their possible contributions to the learning process prior to model training. These metrics leverage the reduction of uncertainty that a feature provides about the target value and a quantification of the influence of the feature on a categorical target measured by evaluating the variability in the feature values across the categories defined by the target. Those values are obtained relative to the dataset - the terms are scaled using the absolute maximum value obtained per feature. The terms of the proposed metric are combined in one of three ways: using either arithmetic, harmonic, or geometric means.

The proposed “Relative Information Availability” metric offers several advantages.

- The proposed metrics combine the strengths of filter methods (speed and simplicity) with the potential to capture interactions through the used aggregation methods (arithmetic, harmonic, geometric means). They provide a more rounded evaluation of feature importance that acknowledges different properties.
- The metrics combine the relative information content of a feature with its relative variance, providing a balanced perspective on what makes a feature important and likely to be relevant in model training. By integrating the quantification of different aspects of the features, the metrics gain expressiveness of the captured complexities.
- While MI can capture non-linear relationships, the F-value can reinforce the importance of features with significant variance effects between classes.
- The metrics provide significant pre-model insights that can guide subsequent model complexity and architecture decisions, potentially improving model performance by focusing on genuinely informative features right from the start. It allows the user to check if the intuition of the information expected from the captured features aligns with what is actually present in the data and available for the AI/ML models to leverage, forming the end-user expectation of the model behavior before any computational effort is spent on training.
- Since these metrics do not rely on a specific model, they can provide a more objective baseline that is not biased by the peculiarities of a particular modeling algorithm.

V. CALCULATING THE PROPOSED NOVEL METRIC: RELATIVE INFORMATION AVAILABILITY

A. DECREASE IN UNCERTAINTY ABOUT THE TARGET VARIABLE

The first term of the equation contained in Eq. 1, known as Mutual Information, quantifies the amount of information that one random variable contains about another random variable. It measures the reduction in uncertainty about one variable given knowledge of the other. It quantifies the dependency between two variables. A high value indicates a strong relationship, where knowing the value of one variable provides significant information about the other.

It is always a non-negative value. When it equals zero, it means that the variables are independent, and there is no information shared between them.

The term is related to statistical measures. For instance, it is related to the Pearson correlation coefficient in that both measure how variables relate, but unlike correlation, the term can capture non-linear relationships and is not limited to linear dependencies. The term does not require assumptions about the distribution of the data (such as normality). $p(x, y)$ is the probability that X takes the value of x and Y takes the value of y simultaneously. It represents how often pairs of values (x,y) occur together. $\frac{p(x,y)}{p(x)p(y)}$ This ratio compares the joint probability of x and y with the product of their individual probabilities. If X and Y are independent, this ratio will be 1. The log function translates probabilities into information content. When this ratio is 1, it indicates independence, and the logarithm of 1 is 0, meaning no mutual information. When the ratio deviates from 1, the logarithm captures the extent of dependence or independence between X and Y.

$$\sum_{y \in Y} \sum_{x \in X} p(x, y) \log \left(\frac{p(x, y)}{p(x)p(y)} \right) \quad (1)$$

B. ANALYSIS OF DIFFERENCES BETWEEN CLASS MEANS

The second term, contained in Eq. 2 is a value quantifying the differences between group variances in a dataset. The goal is to measure the influence of a feature on a target value by examining the variability in the feature values across the categories defined by the target value. The term indicates whether the means of the feature across different categories of the target are different. First, the data for each feature is grouped based on the target categories, then the observations for each feature are split into groups. Two sub-terms are used to produce the value. The nominator measures the variance due to the interaction between the different categories of the target and the feature. Essentially, it is about how much the group means deviate from the overall mean. The denominator measures the variance within each category of the target. It reflects the spread of the feature values within each category. A high value indicates that the between-group variance is significantly larger than the within-group variance, suggesting that the feature effectively differentiates between the categories of the target. In statistics, this term is known as ANOVA F-value.

A high value of the term signifies that the features are considered more predictive and may be prioritized in model training because they show significant differences across the target categories. A low value of the term signifies that the features are less useful in predicting the target variable, as they do not vary much across the target categories.

$$\frac{\frac{1}{k-1} \sum_{i=1}^k n_i (\bar{X}_i - \bar{X})^2}{\frac{1}{N-k} \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2} \quad (2)$$

In Eq.2, k is the number of groups and represents the total number of different categories in the dataset. N is the total number of observations (data points) across all groups.

n_i is the number of observations in a class.

i is the count of data points in the i -th group.

\bar{X} is the mean of all observations across all classes.

\bar{X}_i is the mean of a class i .

C. MAXIMUM ABSOLUTE SCALING (MaxAbs SCALING)

A scaled first term value close to 1 indicates that the feature associated with this value has the strongest association with the target variable compared to all other features being considered. This is because the scaling factor is the maximum absolute value of the term among all features. This is visible in Eq. 3, where each MI for feature f is scaled by the highest obtained MI value. This makes the metric much more readable but also makes all the readings relative to a particular set of features. Scaling highlights the relative importance of features in the context of the information structure of the dataset.

$$\frac{\sum_{y \in Y} \sum_{x \in X_f} p(x, y) \log \left(\frac{p(x, y)}{p(x)p(y)} \right)}{\max_{f \in F} \left(\sum_{y \in Y} \sum_{x \in X_f} p(x, y) \log \left(\frac{p(x, y)}{p(x)p(y)} \right) \right)} \quad (3)$$

Max Absolute Scaling normalizes the values to a $[0,1]$ range (de-facto MaxAbs scales to $[-1,1]$, but both the terms in the metric equation are always positive). A value near one signifies that the information gain (reduction in uncertainty about the target variable) provided by this feature is the highest obtainable in the given dataset. It effectively allows for easier comparison between features. In practical terms, the scaled value suggests that the feature is highly valuable for predicting the target variable, relative to all the other features in that dataset.

The second term, when scaled with MaxAbs, identifies the feature which has the highest between-class variance to within-class variance ratio and assigns it the value of one. This suggests the most significant feature from the variance perspective among the available features, as seen in Eq.4.

$$\frac{\frac{1}{k-1} \sum_{i=1}^k n_i (\bar{X}_{i,f} - \bar{X}_f)^2}{\frac{1}{N-k} \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij,f} - \bar{X}_{i,f})^2}}{\max_{f \in F} \left(\frac{\frac{1}{k-1} \sum_{i=1}^k n_i (\bar{X}_{i,f} - \bar{X}_f)^2}{\frac{1}{N-k} \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij,f} - \bar{X}_{i,f})^2} \right)} \quad (4)$$

With both terms having a range from zero to positive infinity, Maximum Absolute Scaling ensures their comparability, forcing them into the range $[0,1]$.

D. FORMULATING THE METRIC

With the two terms - one expressing the value of a feature in decreasing the uncertainty about the target value, and the other describing the variability in the feature values across the classes, both scaled to be in the same range, there exist a couple of ways of combining them that behave slightly differently.

Arithmetic Mean used in the ARIA (Arithmetic Relative Information Availability) formula in Eq. 5, as shown at the bottom of the next page, offers a straightforward average, a balance between the two terms of the equation, assuming equal significance of both terms.

Harmonic Mean present in the HaRIA formula in Eq. 6, as shown at the bottom of the next page, favors the lower of the two terms, thus it can be particularly useful in scenarios where both scores need to be sufficiently high to consider a feature truly relevant. This can pinpoint features that are informative and likely to be utilized by the model.

The Geometric Mean in the GeRIA formula in Eq. 7, as shown at the bottom of the next page, provides a balance that is less sensitive to extreme values than the ARIA, which can be useful when combining terms that happen to have different distributions. Naturally, this makes it similar to the ARIA in this context, as with MaxAbs Scaling employed in the pipeline the possibility of encountering extreme values is minimal. It is included in the evaluation for the sake of completeness.

VI. EXPERIMENTAL SETUP

A. DATASETS

The proposed metrics were tested on 7 different datasets with ranging properties, including ML classics like Wine and Iris, but also e.g. a real-world cybersecurity Netflow dataset 'ToN'. The full list of employed datasets is collected in Tab.1.

All the experiments are completed in the framework of a 10-fold cross-validation, and the reported metrics are averages over all 10 folds.

In each fold, the data was preprocessed with sci-kit learns StandardScaler.

The 9/10th training sets are used to train 3 classifiers: RF, ET, and an ANN (using TensorFlow). The network has 2 hidden layers, both using the ReLU activation function, 64 neurons on the first layer, 32 on the second layer, softmax output, and categorical cross-entropy as the loss function. The optimizer was ADAM. The choice of hyperparameters was grounded in previous research and could be subject to further refinements. However, since the aim of the research was not to present the best classifier but to propose a PMI metric, the most optimal hyperparameter setup for all the tested datasets was not pursued.

The purpose of training the models is to evaluate how they assign importance to different features learning to classify.

Various methods are employed to compare feature importance both pre- and post- model training. Mutual Information and ANOVA F-values provide statistical measures of feature relevance. Model-specific feature importance scores are

TABLE 1. Detailed overview of datasets used in the analysis with specific file references.

Dataset Identifier	Dataset Name	# Features	# Samples	# Classes	File Name
[40]	Iris dataset	4	150	3	iris_20240403-190919.csv
[41]	Wine dataset	13	178	3	wine_20240329-201912.csv
[42]	Diabetes dataset	10	442	214 (treated as classification)	diabetes_20240408-132635.csv
[43]	Dry Bean dataset	16	13611	7	drybean_20240328-193526.csv
[44]	Rice dataset	7	3810	2	rice_20240330-022241.csv
[45]	Glass dataset	9	214	6	glass_20240330-141256.csv
[46] [47]	NetFlow ToN IoT dataset	8	1379274	10	Ton_20240410-144624.csv

extracted directly from ensemble tree models (RandomForestClassifier and ExtraTreesClassifier). SHAP (SHapley Additive exPlanations) values are computed to explain the output of the models, particularly the contribution of each feature to the classification.

To make the comparison more evident, scores scaled with maximum absolute scaling are provided. This allows for a direct comparison across different metrics and models.

The relationships among the gathered results are showcased using Parallel Coordinate Plots and Heatmaps.

VII. RESULTS

In the parallel coordinate plots, each vertical line represents a different metric for assessing the features. From left to right, the labels represent: Scaled Mutual Information (Scaled MI), Scaled F (Scaled ANOVA F-values), Scaled Random Forest Importance (RF Importance Scaled), The MaxAbs scaled aggregated SHAP value (RF Scaled Agg Abs SHAP), Feature Importance and scaled aggregated SHAP for the Extra Trees Classifier, scaled aggregated SHAP for the ANN, and then the three proposed metrics - the HaRIA, the ARIA and the GeRIA. Each line represents a different feature, marked by a distinct color. The way the lines cut across the plot indicates how each feature ranks according to the different metrics. This plot is valuable for understanding how different metrics assess the feature value or contribution to the

classification, underlining either consistency or discrepancies between established methods and the proposed metric.

While the parallel coordinate plots provide a high-level overview, the heatmap of scaled values is especially informative. In this case, all the values are in the range between zero and one. This allows for a direct comparison between the measures, one that is even easier to understand thanks to the seaborn's [48] coloring capabilities, which allow one to set blue colors to values nearing zero, and red colors to values near one.

A. ANALYSIS OF RESULTS

In Fig. 2, a noticeable pattern present in the Diabetes dataset is the high variability of some values across different metrics. These features score very high in certain metrics (like RF Importance-Scaled) but lower in others (like Scaled MI or Scaled F). The three proposed metrics (HaRIA, ARIA, GeRIA) show a convergence trend where the extremes of importance seen in earlier metrics are moderated. This indicates that the proposed metrics tend to provide a more balanced view that potentially averages out the extremities observed in individual assessments. The most important aspect of the proposed metrics (the three values on the right), which is very visible in the case of this dataset, is that they correlate strongly with the model-based impor-

$$ARIA(X_f, Y) = \frac{\frac{\sum_{y \in Y} \sum_{x \in X_f} p(x, y) \log\left(\frac{p(x, y)}{p(x)p(y)}\right)}{\max_{f \in F} \left(\sum_{y \in Y} \sum_{x \in X_f} p(x, y) \log\left(\frac{p(x, y)}{p(x)p(y)}\right)\right)} + \frac{\frac{\sum_{i=1}^k n_i (\bar{X}_{i,f} - \bar{X}_f)^2}{k-1}}{\max_{f \in F} \left(\frac{\sum_{i=1}^k n_i (\bar{X}_{i,f} - \bar{X}_f)^2}{k-1}\right)}}{2} \tag{5}$$

$$HaRIA(X_f, Y) = \frac{2 \left(\frac{\sum_{y \in Y} \sum_{x \in X_f} p(x, y) \log\left(\frac{p(x, y)}{p(x)p(y)}\right)}{\max_{f \in F} \left(\sum_{y \in Y} \sum_{x \in X_f} p(x, y) \log\left(\frac{p(x, y)}{p(x)p(y)}\right)\right)} \right) \left(\frac{\sum_{i=1}^k n_i (\bar{X}_{i,f} - \bar{X}_f)^2}{k-1} \right)}{\left(\frac{\sum_{y \in Y} \sum_{x \in X_f} p(x, y) \log\left(\frac{p(x, y)}{p(x)p(y)}\right)}{\max_{f \in F} \left(\sum_{y \in Y} \sum_{x \in X_f} p(x, y) \log\left(\frac{p(x, y)}{p(x)p(y)}\right)\right)} \right) + \left(\frac{\sum_{i=1}^k n_i (\bar{X}_{i,f} - \bar{X}_f)^2}{k-1} \right)}{\left(\frac{\sum_{i=1}^k n_i (\bar{X}_{i,f} - \bar{X}_f)^2}{k-1} \right)} \tag{6}$$

$$GeRIA(X_f, Y) = \sqrt{\frac{\sum_{y \in Y} \sum_{x \in X_f} p(x, y) \log\left(\frac{p(x, y)}{p(x)p(y)}\right)}{\max_{f \in F} \left(\sum_{y \in Y} \sum_{x \in X_f} p(x, y) \log\left(\frac{p(x, y)}{p(x)p(y)}\right)\right)} \cdot \frac{\sum_{i=1}^k n_i (\bar{X}_{i,f} - \bar{X}_f)^2}{k-1}}{\max_{f \in F} \left(\frac{\sum_{i=1}^k n_i (\bar{X}_{i,f} - \bar{X}_f)^2}{k-1}\right)} \tag{7}$$

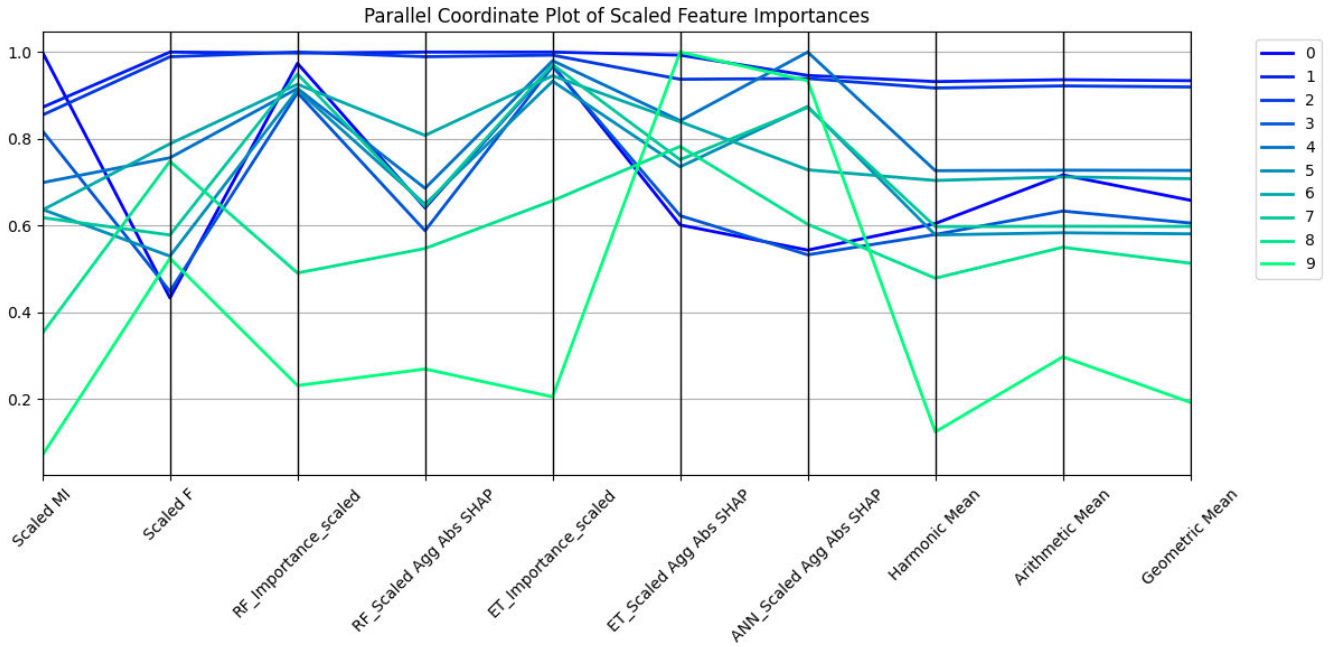


FIGURE 2. The diabetes dataset parallel coordinate plot.

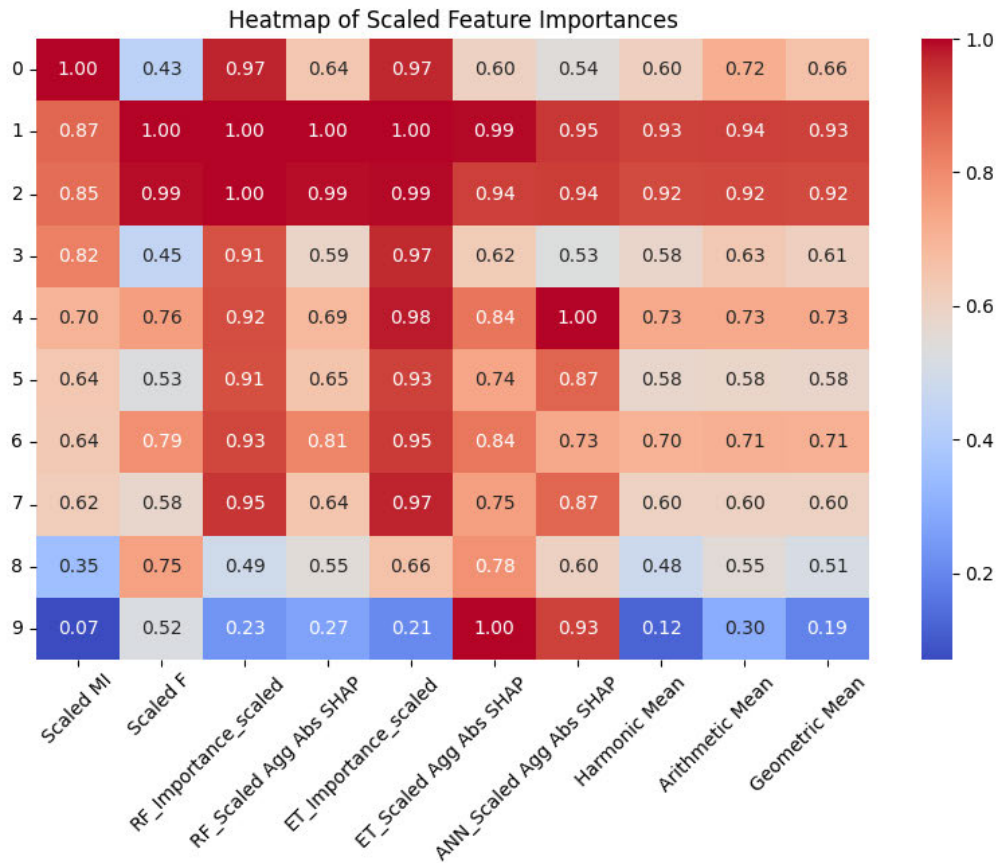


FIGURE 3. The diabetes dataset heatmap.

tances (RF-importance, ET-importance), providing similar importance scores, without the need of any model training.

The heatmap in Fig. 3 shows that certain metrics, like Scaled MI, have lower correlations with model-specific metrics (SHAP values, RF importance, ET importance).

This could indicate a discrepancy in what is considered important by filter methods versus model-based evaluations. It is most evident in features 0, 1, 2 and 6, where Scaled MI of feature 0 is the highest among the features, but it is not the most important feature according to the RF and ET importance metrics. The features 1 and 2 are scored very high by all metrics except Mutual Information. This could lead to a conclusion that MI is not that important for feature selection; however, features 6 and 7 demonstrate that a combination of sufficient MI and variance, as captured by the F-value, can still render these features important in model-specific assessments, despite not being standout in individual statistical evaluations. This interplay suggests that both MI and variance contribute enough unique value that, when combined, enhances the predictive power of these features in the context of the employed models, precisely the kind of interplay the proposed metrics attempt to capture and express.

The parallel coordinate plot and heatmap for the ToN NetFlow collection dataset (Fig. 4 and Fig. 5), a real-world dataset used in ML-based Network Intrusion Detection (NIDS) research, displays the other side of the coin. The feature marked '0' possesses the highest MI of all the features, but its F-value is relatively lower, much lower than the top feature - '4'. However, the high MI values of features 1, 2 and 3 are not usable for the ML models since the F-value is way too low. The proposed metrics moderate the extremes seen in single metrics, providing a more reliable indicator of overall feature importance. It is worth pointing out that the ANN, as indicated by the SHAP value, puts major importance on feature 4, displaying a different decision process to the tree-based models. This is expressed in the proposed metrics, which indicate that both feature 0 and feature 4 display potential for model learning. It is important to bear in mind that while the values of RIA might not pinpoint the exact values reported by posthoc importance scores or SHAP values, they provide usable and informative estimates without any costly model training. The proposed metrics also capture the fact that features 1, 2, and 3 lose the ability to convey the decrease of uncertainty expressed by MI due to very low variance.

In the Wine dataset, Fig.6 and Fig.7, there are notable correlations among model-based metrics (RF, ET, SHAP values) and the proposed RIA metrics, indicating agreement. As in the previous instances, the lower variance expressed by the F-value mitigates the usefulness of features which might otherwise be very strong, as indicated by MI. However, feature 2 displays the opposite behavior, where the relatively low f-value is offset by the high information content (MI), with most of the model-based metrics and the proposed metrics clearly indicating the importance of that feature. Features 0 and 4 maintain high importance across almost all metrics, suggesting they are crucial for models in this dataset. This is captured and expressed by the proposed metrics.

In the Drybean dataset, as shown in Fig.8 and Fig.9 the model-based metrics are aligned with the proposed metrics in the top-performing features, suggesting that the proposed

metrics are valuable for indicating possible feature contributions. An interesting discrepancy between the tree-based models' importance metrics and the calculated SHAP values is visible in those figures - where the Importance Scores are much higher than the calculated SHAP contributions of features 3, 4, 5, 6, and 7. This could arise from several underlying reasons. RF and ET calculate feature importance based on how effectively the features contribute to the model's accuracy. The metrics rely on the average decrease in impurity (using Gini) brought about by splits over each feature across all trees in the forest. If a feature consistently appears in splits that help in substantially reducing impurity, its importance score will be high. SHAP values, on the other hand, measure the contribution of each feature to the prediction of each individual sample, considering all possible combinations of features. SHAP values thus offer a value contribution to the prediction of a specific instance, averaged over many possible coalitions of feature sets. In tree-based models, features that interact complexly with other features might result in higher importance scores because they significantly affect model performance when used in splits. However, these features might not show high SHAP values if the predictive power they contribute is diffused when averaged across all possible feature subsets, especially if their effect is highly context-dependent on the presence or absence of other features. If multiple features carry similar information, tree-based importance might still attribute high scores to all such features if they individually help in making good predictions in different parts of the tree structures. SHAP values might spread the attribution more thinly among these features because they account for the presence of correlated or redundant information.

The proposed metrics not only smooth out extreme values but also highlight features that consistently appear important across different methodologies, thereby potentially increasing their reliability.

Analyzing the parallel coordinate plot and heatmap for the Iris dataset in Fig.10 and Fig.11, it is apparent that features 0 and 1 maintain a higher importance across the metrics, including the proposed ones, indicating their important role in classification. The heatmap shows clearly that the post-hoc metrics agree with the proposed pre-model metrics as to the feature importance.

In Fig.12 and Fig.13, the feature metrics for the Rice dataset are gathered. Similar to the previous evaluations across different datasets, the figures reveal that while there is some variance in how the proposed metrics relate to model-based metrics, they generally follow a similar trend. This indicates an alignment of the proposed metrics with the model-based interpretations, reinforcing their practical applicability understanding the nature of the dataset. One observation that needs to be addressed in the Rice dataset is that the ANN chose different features as important, according to accumulated SHAP values. There is a number of reasons why this could happen; for example, features that do not perform well individually in tree splits could be vital in

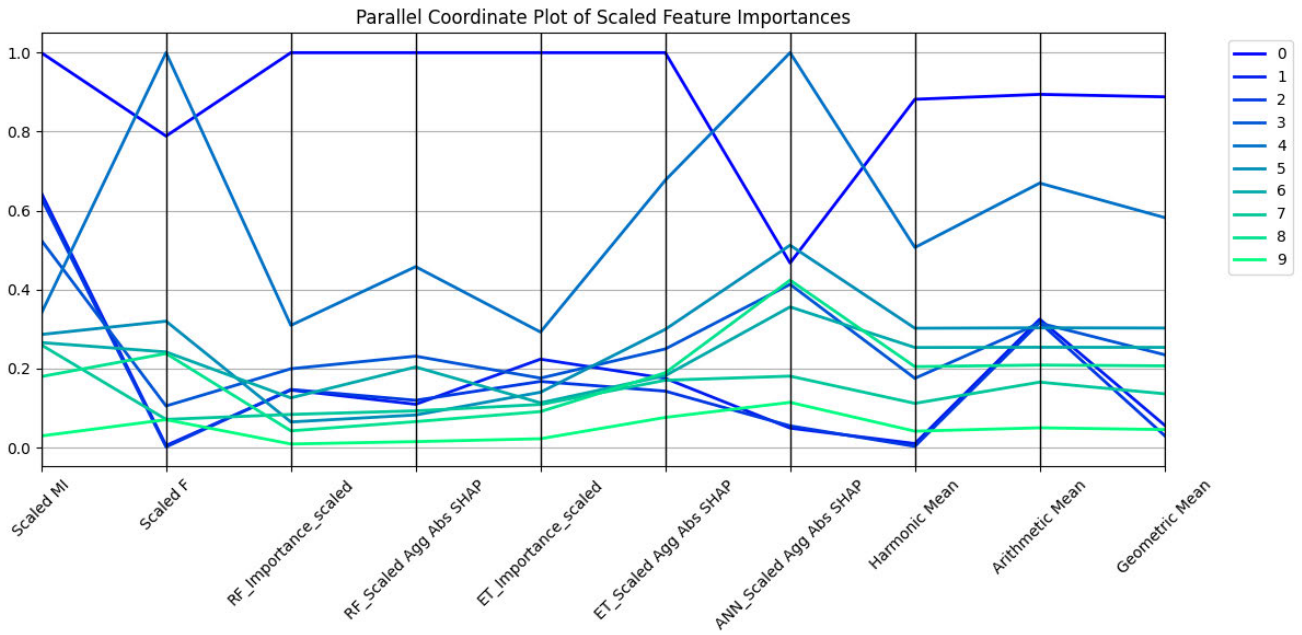


FIGURE 4. The ToN NetFlow collection dataset parallel coordinate plot.

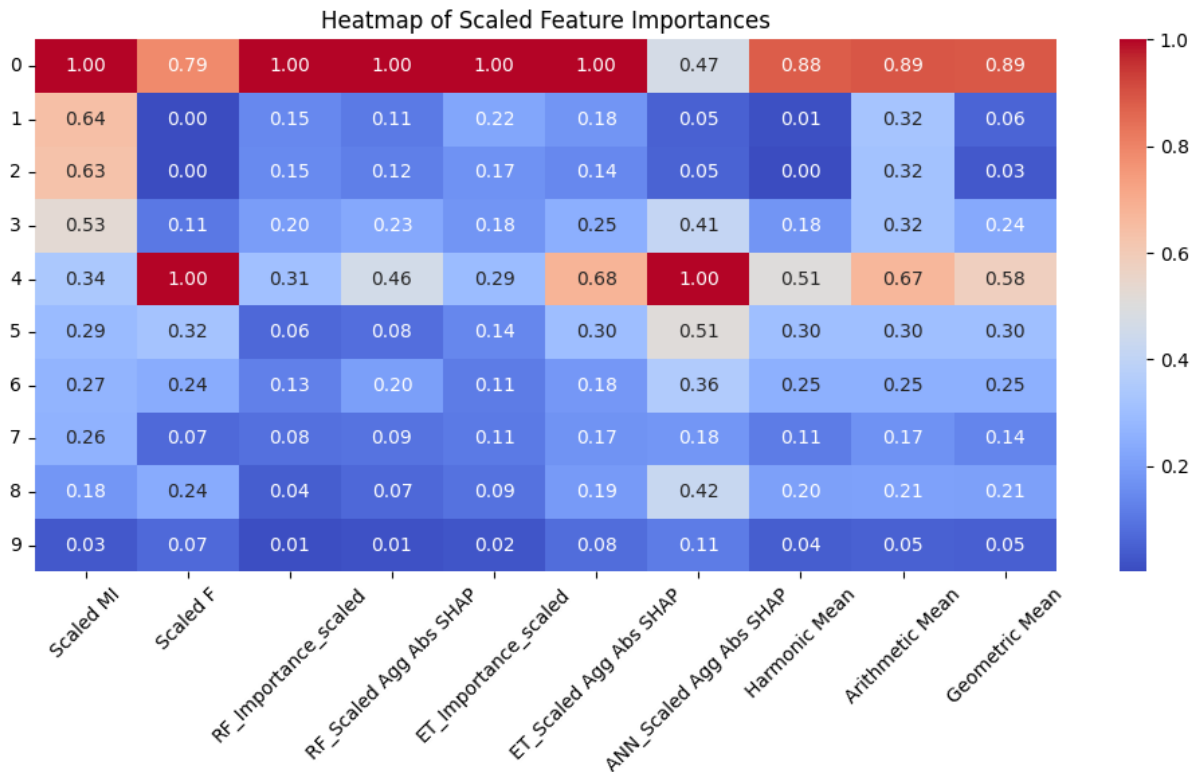


FIGURE 5. The ToN NetFlow collection dataset heatmap.

an ANN if they interact synergistically with other features to affect the output. SHAP values for ANNs are computed based on the contribution of each feature to the output across potentially complex nonlinear transformations, reflecting a cumulative effect of features throughout all the layers.

In contrast, tree-based SHAP calculations directly relate to the decrease in model error or impurity, typically measured in a more straightforward manner. It is important to point out that, according to the experiments presented in this paper, as showcased in the figures, this happens when MI scores and

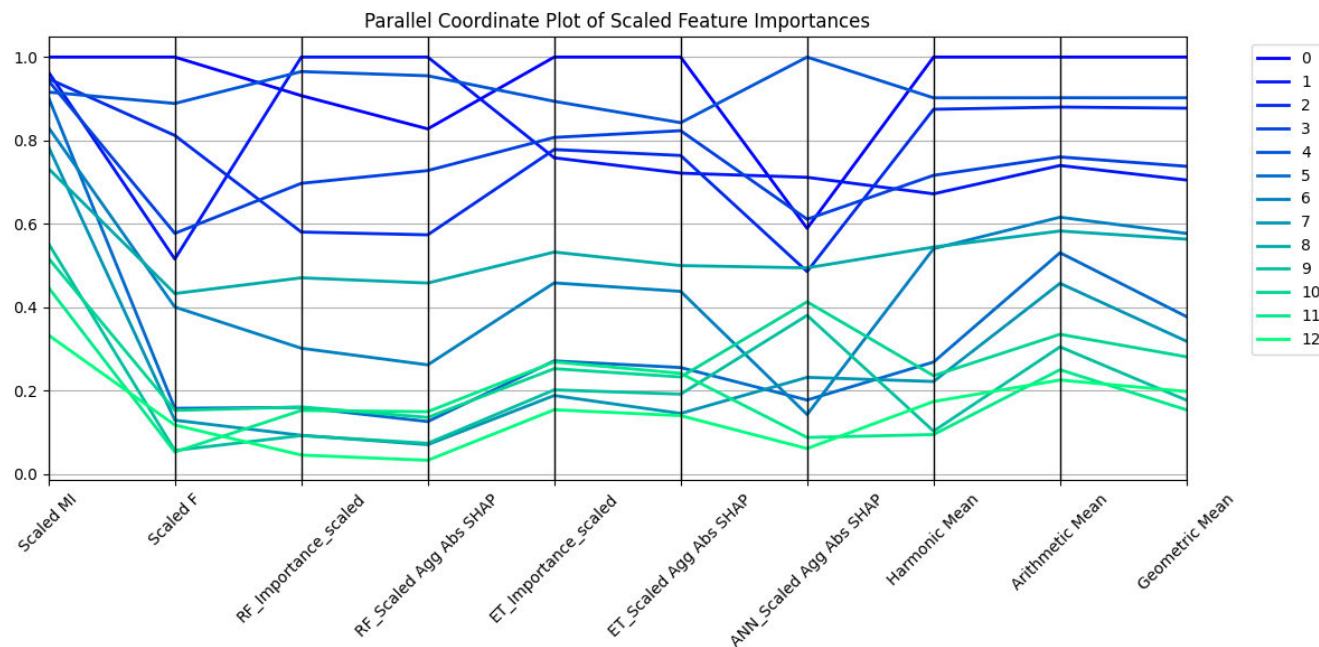


FIGURE 6. The wine dataset parallel coordinate plot.

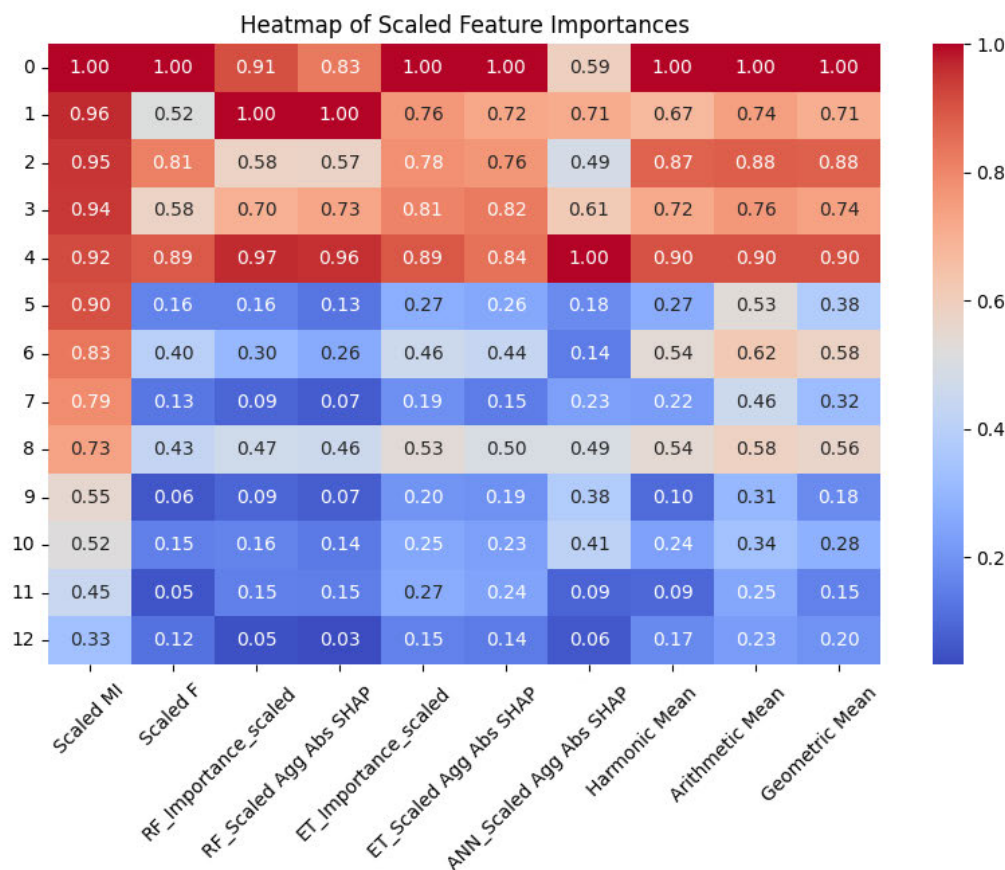


FIGURE 7. The wine dataset heatmap.

F-values are relatively high in multiple features, allowing for the different ways the algorithms learn to be brought to light.

In Fig.14 and Fig.15, it is evident that features 4 and 5 are consistently identified as important across all

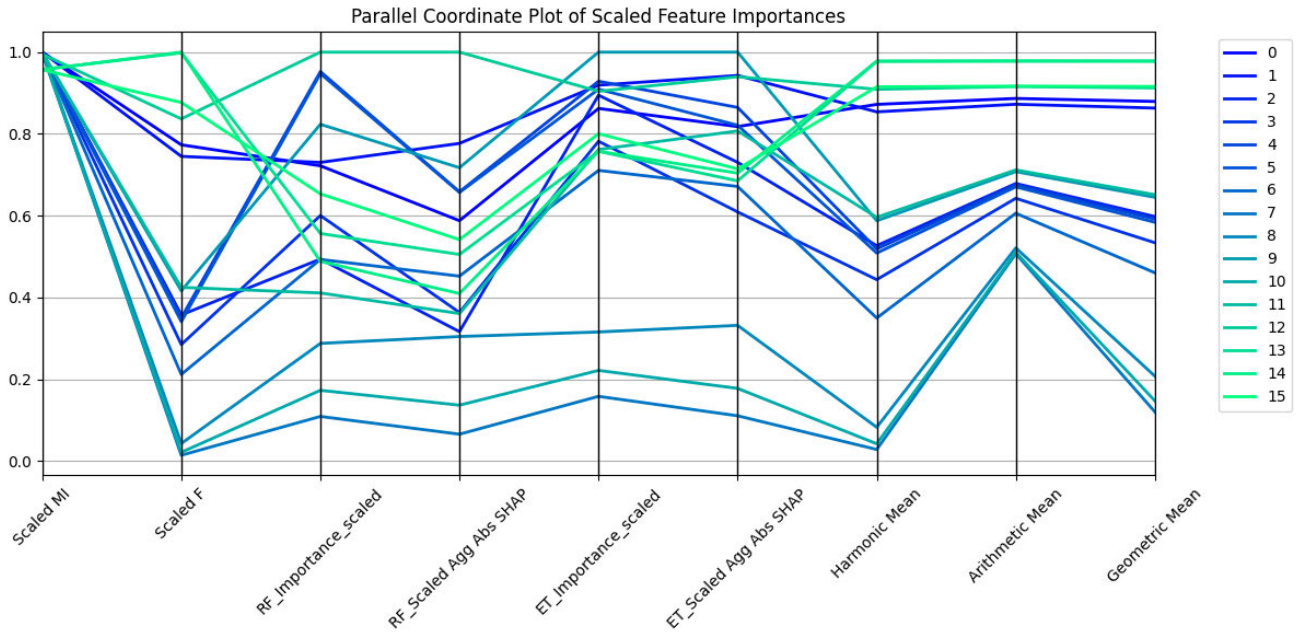


FIGURE 8. The Drybean dataset parallel coordinate plot.

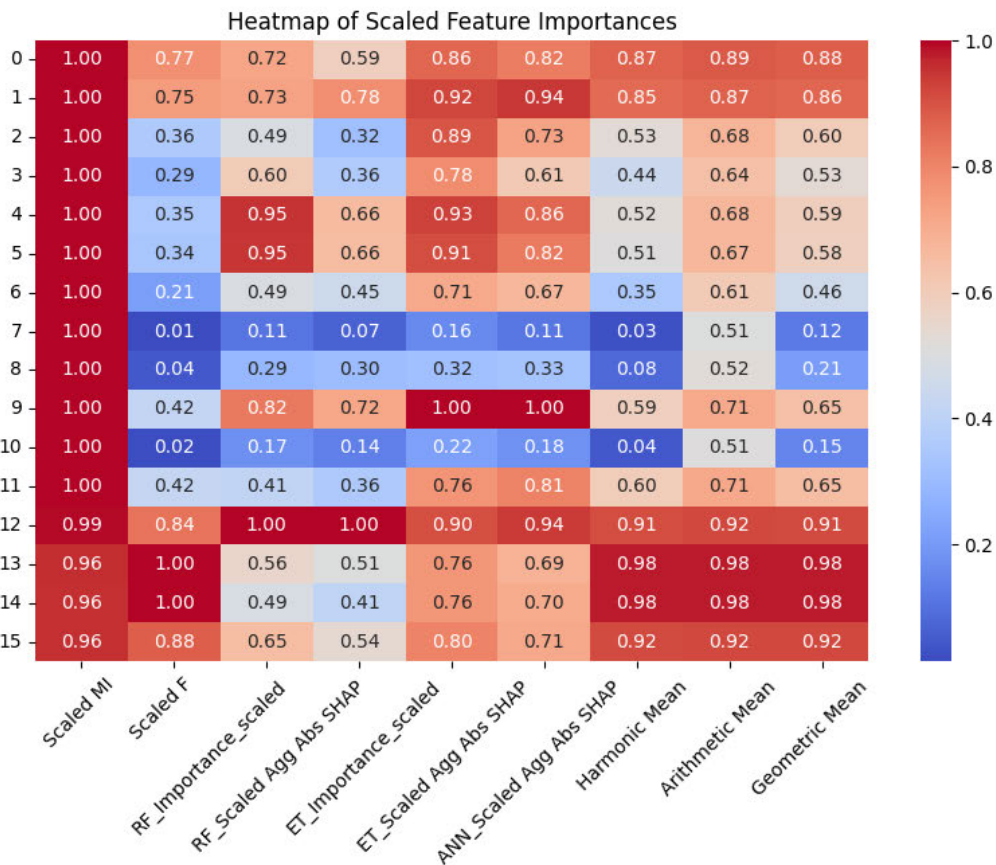


FIGURE 9. The Drybean dataset heatmap.

methods, reinforcing their role as key predictors in the dataset. The proposed metrics capture well the essential aspects of those features. The outlier of this dataset is

feature 0, which provides very little variance, but the highest MI. Thus, it becomes a relatively important feature as valued by RF Importance, RF SHAP, ET Importance

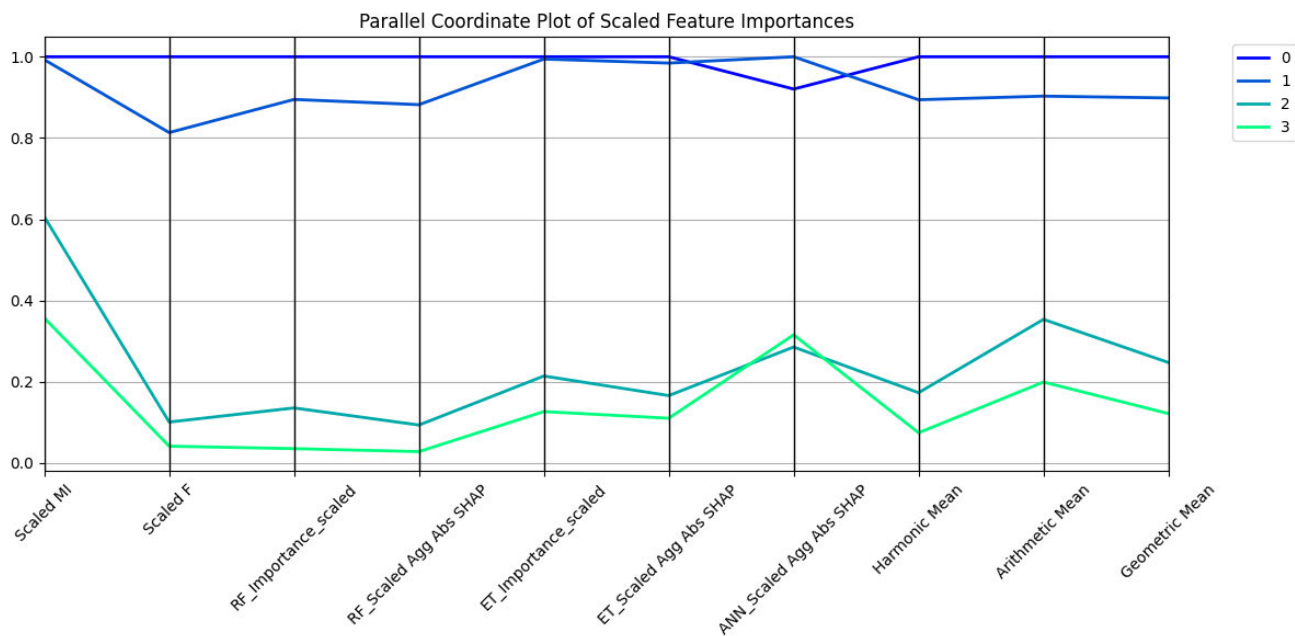


FIGURE 10. The Iris dataset parallel coordinate plot.

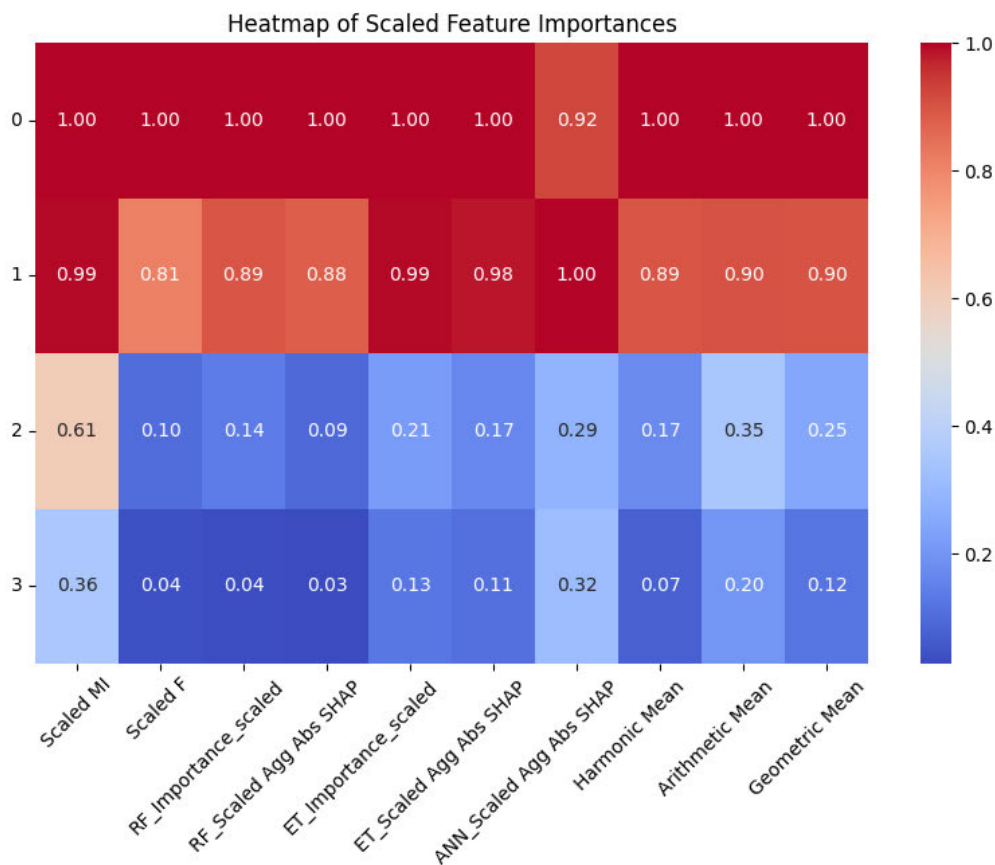


FIGURE 11. The Iris dataset heatmap.

and ANN SHAP. The proposed metrics undervalue this feature in this dataset, except for the straightforward ARIA.

B. STATISTICAL ANALYSIS

The Table 2 shows the Pearson correlation coefficients between various feature importance measurements derived

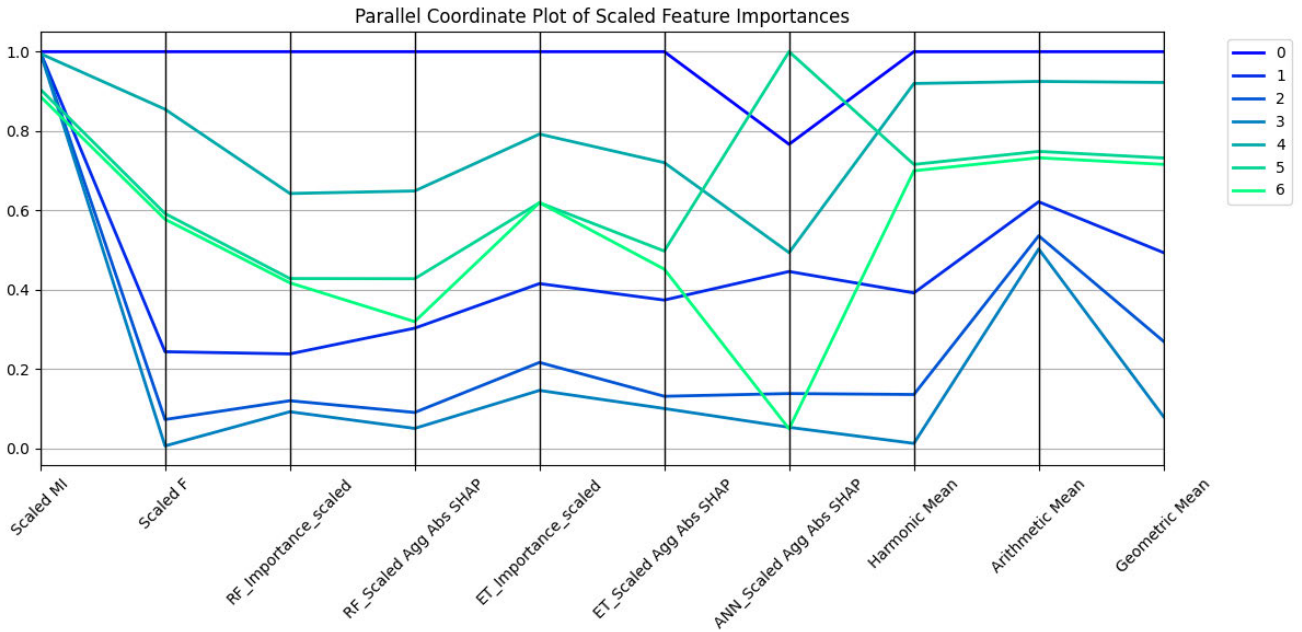


FIGURE 12. The rice dataset parallel coordinate plot.

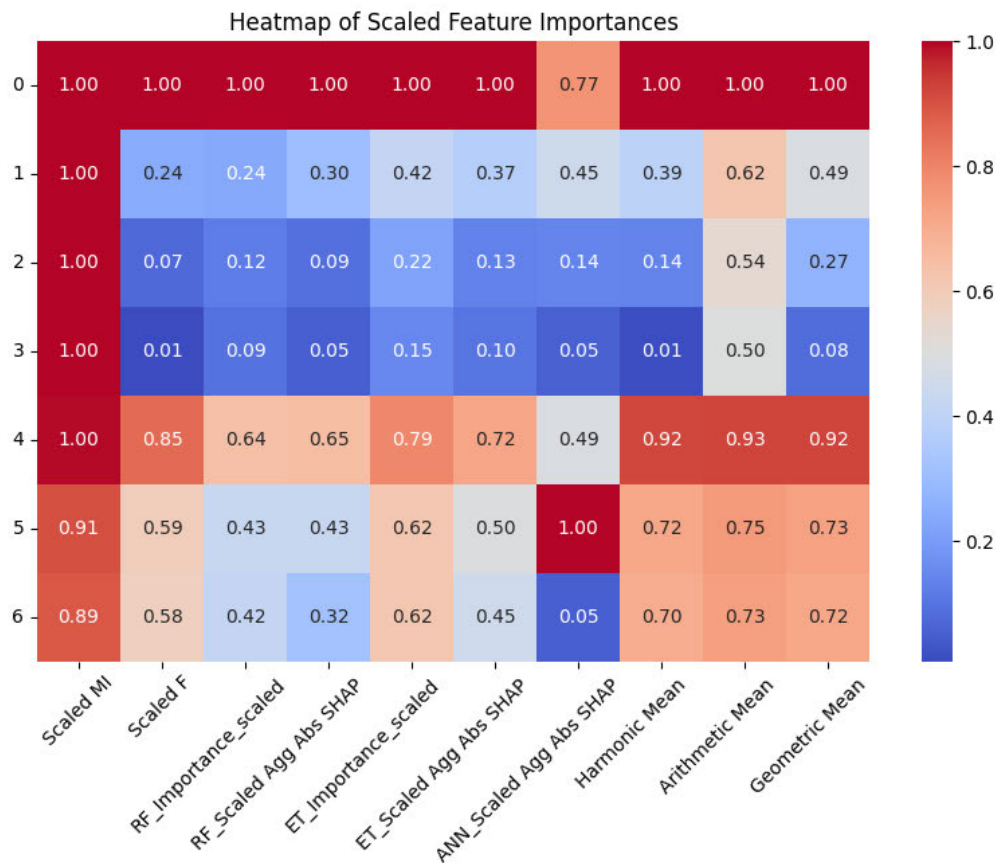


FIGURE 13. The rice dataset heatmap.

from different ML models and the three proposed metrics (HaRIA, ARIA, GeRIA) across multiple datasets (ToN_IoT, Wine, Iris, Diabetes, Rice, Glass, and Drybean). These

correlations indicate how closely the proposed metrics reflect the importance assigned by the model mechanisms or SHAP values, which represent the contribution of each feature

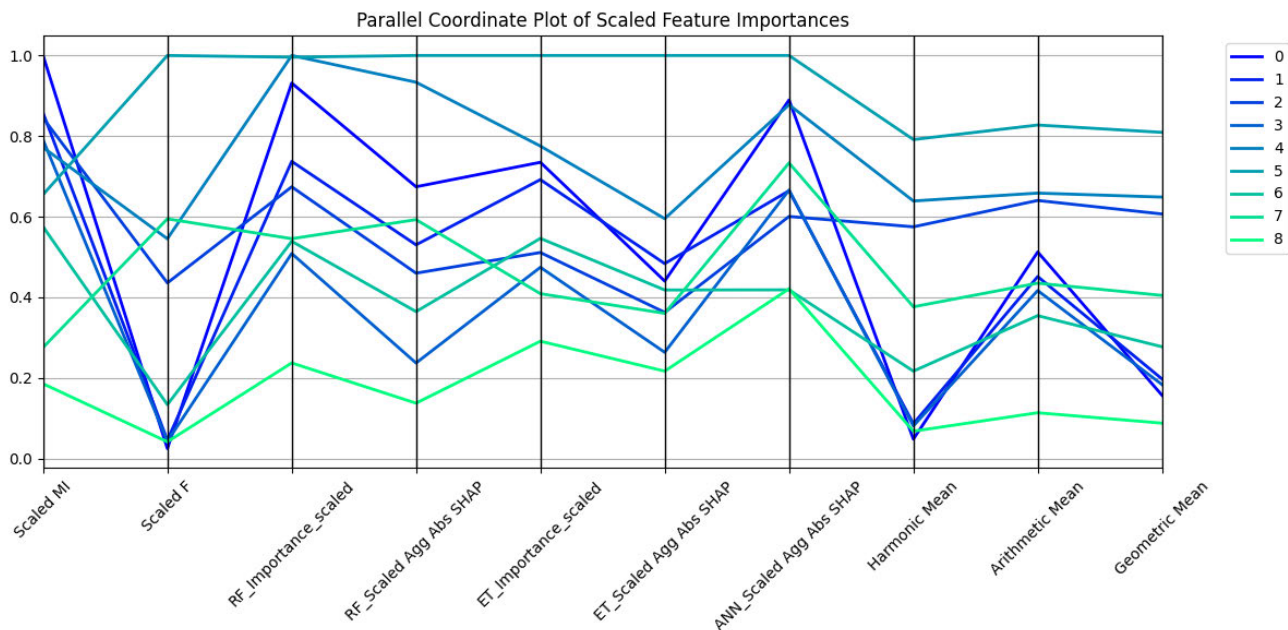


FIGURE 14. The glass dataset parallel coordinate plot.

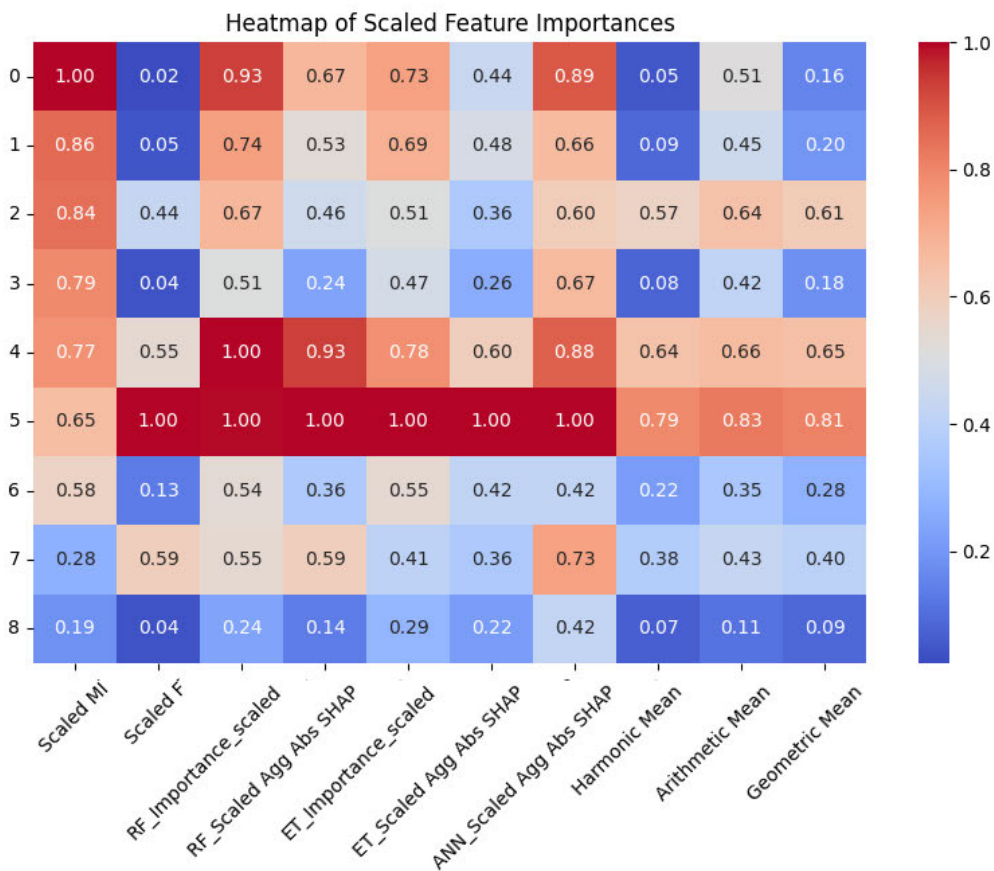


FIGURE 15. The glass dataset heatmap.

to the model’s prediction. Higher correlations suggest that the proposed metric effectively captures the importance that the models attribute to the features. The Iris dataset

shows extremely high correlations for all three proposed metrics, especially with features’ importance measured by the RF model. Correlations are nearly perfect, close to 1.

TABLE 2. Pearson correlation of feature importances measured in different ways with the proposed metrics, across different datasets.

Feature	Correlation with HaRIA	Correlation with ARIA	Correlation with GeRIA	Dataset
RF_Importance_scaled	0.87444	0.90712	0.87707	ToN
RF_Scaled Agg Abs SHAP	0.92545	0.93913	0.93472	ToN
ET_Importance_scaled	0.86113	0.90014	0.85972	ToN
ET_Scaled Agg Abs SHAP	0.96045	0.96453	0.97619	ToN
ANN_Scaled Agg Abs SHAP	0.66161	0.57921	0.69730	ToN
RF_Importance_scaled	0.90385	0.88366	0.90164	Wine
RF_Scaled Agg Abs SHAP	0.89099	0.86769	0.88768	Wine
ET_Importance_scaled	0.97213	0.94811	0.96811	Wine
ET_Scaled Agg Abs SHAP	0.96841	0.94348	0.96401	Wine
ANN_Scaled Agg Abs SHAP	0.75080	0.73585	0.74796	Wine
RF_Importance_scaled	0.99999	0.99687	0.99942	Iris
RF_Scaled Agg Abs SHAP	0.99948	0.99424	0.99803	Iris
ET_Importance_scaled	0.99630	0.99185	0.99530	Iris
ET_Scaled Agg Abs SHAP	0.99627	0.98975	0.99440	Iris
ANN_Scaled Agg Abs SHAP	0.97791	0.96465	0.97340	Iris
RF_Importance_scaled	0.85429	0.81331	0.84491	Diabetes
RF_Scaled Agg Abs SHAP	0.97634	0.96515	0.97482	Diabetes
ET_Importance_scaled	0.85597	0.80228	0.84351	Diabetes
ET_Scaled Agg Abs SHAP	0.08882	0.10342	0.08298	Diabetes
ANN_Scaled Agg Abs SHAP	0.19900	0.14118	0.16582	Diabetes
RF_Importance_scaled	0.91715	0.97498	0.91272	Rice
RF_Scaled Agg Abs SHAP	0.90529	0.96790	0.90759	Rice
ET_Importance_scaled	0.98419	0.98702	0.98160	Rice
ET_Scaled Agg Abs SHAP	0.94445	0.98375	0.94470	Rice
ANN_Scaled Agg Abs SHAP	0.63238	0.60907	0.63846	Rice
RF_Importance_scaled	0.69840	0.79211	0.72810	Glass
RF_Scaled Agg Abs SHAP	0.79213	0.81322	0.81130	Glass
ET_Importance_scaled	0.70721	0.81783	0.70117	Glass
ET_Scaled Agg Abs SHAP	0.72324	0.84801	0.75600	Glass
ANN_Scaled Agg Abs SHAP	0.51669	0.80122	0.62040	Glass
RF_Importance_scaled	0.55572	0.87330	0.62040	Drybean
RF_Scaled Agg Abs SHAP	0.72617	0.85478	0.75994	Drybean
ET_Importance_scaled	0.53811	0.82810	0.60219	Drybean
ET_Scaled Agg Abs SHAP	0.71731	0.80643	0.75229	Drybean
ANN_Scaled Agg Abs SHAP	0.51669	0.80122	0.56820	Drybean

This suggests that for the Iris dataset, all three proposed metrics (HaRIA, ARIA, GeRIA) are very effective in aligning with the models' view on feature importance. For the ToN and Wine datasets, the proposed metrics also show strong correlations with the traditional and SHAP-based feature importance scores, indicating good alignment across these metrics. Rice, Glass, and Drybean datasets show generally good correlations, particularly between ARIA and the model-based importance measures. The results suggest that the ARIA metric performs consistently well across different datasets, and when it is surpassed by HaRIA, it still maintains high performance. The Diabetes dataset presents a more varied range of correlations, with some metrics like the ET Scaled Agg Abs SHAP showing very low correlation scores, especially with GeRIA. This might be the result of the Diabetes dataset being a regression dataset, which was treated as a classification dataset to evaluate performance with a high number of classes but a low number of instances in those classes. The ANN Aggregated Max-Abs Scaled SHAP typically shows lower correlations compared to other model-based importances. This might be due to the different nature of feature interactions or different importance attribution in ANNs compared to tree-based models; however, without a true baseline of how feature importance and contribution are

calculated in ANNs, it is not possible to clearly establish whether the proposed metrics miss the mark and the ANN models behave very differently, or if what SHAP shows as feature importance actually correlates with what happens in the model. This is definitely an area for future research. This highlights the challenges in creating a universally effective metric. Table 2 suggests that while the proposed metrics can approximate feature importance very well for some models, they may not capture all nuances.

VIII. CONCLUSION

The proposed metrics - HaRIA, ARIA and GeRIA - facilitate a comprehensive understanding of which features could be utilized in learning before engaging in costly model training. By incorporating MI and F-values, scaled with MaxAbs Scaling, these metrics effectively quantify the potential contribution of each feature to model performance. The metrics capture both the non-linear dependencies and variance effects among features. The proposed RIA metrics provide an assessment of feature importance free of model peculiarities, helping to guide both the initial stages of model development and further analysis of whether the decision-making process of the model is aligned with what is present in the data. By providing insights into feature relevance before any model

training, these metrics allow practitioners to make informed decisions about model architecture and complexity.

The validation of the proposed RIA metrics across multiple datasets showcases the applicability of the proposed metrics. The correlation of RIA metrics with traditional feature importance measures and SHAP values in model training highlights their capability to approximate the importance assigned by complex models effectively. The consistency observed across various datasets in aligning with model-based metrics underscores the potential of RIA metrics. Essentially, the proposed metrics give similar insights but are significantly less expensive.

Answer to RQ1: Can potential feature utilization in an ML model be reliably assessed before any model training? Yes, the proposed RIA metrics—HaRIA, ARIA, and GeRIA—demonstrate the ability to reliably assess potential feature utilization before any model training by quantifying the information content and variance contributions of each feature.

Answer to RQ2: Can pre-model interpretability metrics provide a reliable estimation of feature importance that aligns with traditional post-model training evaluations? Yes, the validation of RIA metrics shows strong correlations with traditional post-model training evaluations, such as feature importance scores from Random Forests or Extra Trees, indicating that these pre-model metrics can reliably estimate feature importance.

Answer to RQ3: Can the proposed pre-model interpretability metrics provide a reliable estimation of what features will be utilized by the trained model, that aligns with traditional post-hoc xAI evaluations? Yes, the proposed metrics align well with post-hoc xAI evaluations like SHAP values across various datasets, thereby providing a reliable estimation of the features that will be utilized by the trained model.

The findings suggest that the RIA metrics can be instrumental in the early stages of model development, providing a reliable estimation of feature importance that aligns well with post-model training evaluations.

However, some discrepancies noted, particularly with complex models like the ANNs, suggest areas for future exploration. The lower correlations between RIA metrics and ANN SHAP values indicate potential differences in how features are utilized by ANNs versus tree-based models, or how ANN SHAP values are calculated vs. the SHAP values for tree-based models. This discrepancy underscores the need for further research.

Additionally, while the metrics perform well across various datasets, their performance in highly specialized or unconventional datasets requires further investigation.

HaRIA, ARIA, and GeRIA not only improve the understanding of the importance of certain features relative to other features in a dataset but also provide a measure of interpretability before any model training is performed. Their development and validation are a promising direction and hopefully will bring more research in this field in ML, which

one day could result in providing a solid ground truth for the current xAI methods.

Future work will incorporate the notion of synergistic relationships between features. It is possible for there to be synergy between two features which is not captured by individual MI, and therefore the RIA metrics. One must consider a situation where X and Z are two binary variables, and Y is a third binary variable that depends on both X and Z in an exclusive or (XOR) relationship. X and Z can both be 0 or 1. If X and Z are the same, $Y=0$. If X and Z are different, $Y=1$. Knowing X alone provides no information about Y because Y can be either 0 or 1 regardless of X . Knowing Z alone also provides no information about Y for the same reason. However, knowing X and Z allows one to determine Y perfectly. Information from knowing X and Z together is greater from knowing X and Z separately. In future work, this effect will be investigated and incorporated into the PMI metrics.

REFERENCES

- [1] D. Sharma and N. Kumar, "A review on machine learning algorithms, tasks and applications," *Int. J. Adv. Res. Comput. Eng. Technol. (IJARCET)*, vol. 6, no. 10, pp. 1323–2278, 2017.
- [2] Z. Goldfeld and Y. Polyanskiy, "The information bottleneck problem and its applications in machine learning," *IEEE J. Sel. Areas Inf. Theory*, vol. 1, no. 1, pp. 19–38, May 2020.
- [3] I. El Naqa and M. J. Murphy, *What Is Machine Learning?*. Cham, Switzerland: Springer, 2015, pp. 3–11.
- [4] A. N. Tikhonov, "Solution of incorrectly formulated problems and the regularization method," *Sov Dok.*, vol. 4, pp. 1035–1038, Jul. 1963.
- [5] M. Stone, "Cross-validated choice and assessment of statistical predictions," *J. Roy. Stat. Soc. Ser. B, Stat. Methodol.*, vol. 36, no. 2, pp. 111–133, Jan. 1974.
- [6] M. Choraś, M. Pawlicki, D. Puchalski, and R. Kozik, "Machine learning—the results are not the only thing that matters! what about security, explainability and fairness?" in *Proc. 20th Int. Conf.*, 2020, pp. 615–628.
- [7] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics*, vol. 8, no. 8, p. 832, Jul. 2019.
- [8] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artif. Intell.*, vol. 267, pp. 1–38, Feb. 2019.
- [9] M. Szczepański, M. Choraś, M. Pawlicki, and A. Pawlicka, "The methods and approaches of explainable artificial intelligence," in *Proc. Int. Conf. Comput. Sci.*, 2021, pp. 3–17.
- [10] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.
- [11] F. Z. Janane, T. Ouaderhman, and H. Chamlal, "A filter feature selection for high-dimensional data," *J. Algorithms Comput. Technol.*, vol. 17, Jan. 2023, Art. no. 17483026231184171.
- [12] K. Hussain, N. Neggaz, W. Zhu, and E. H. Houssein, "An efficient hybrid sine-cosine Harris hawks optimization for low and high-dimensional feature selection," *Expert Syst. Appl.*, vol. 176, Aug. 2021, Art. no. 114778.
- [13] L. Jundong, "Feature selection: A data perspective," *Comput. Surveys*, vol. 50, p. 6, Jun. 2017.
- [14] J. Wang, J. Xu, C. Zhao, Y. Peng, and H. Wang, "An ensemble feature selection method for high-dimensional data based on sort aggregation," *Syst. Sci. Control Eng.*, vol. 7, no. 2, pp. 32–39, Nov. 2019.
- [15] C. Fahy and S. Yang, "Dynamic feature selection for clustering high dimensional data streams," *IEEE Access*, vol. 7, pp. 127128–127140, 2019.
- [16] M. García-Torres, F. Gómez-Vela, B. Melián-Batista, and J. M. Moreno-Vega, "High-dimensional feature selection via feature grouping: A variable neighborhood search approach," *Inf. Sci.*, vol. 326, pp. 102–118, Jan. 2016.

- [17] H. M. Abdulwahab, S. Ajitha, and M. A. N. Saif, "Feature selection techniques in the context of big data: Taxonomy and analysis," *Int. J. Speech Technol.*, vol. 52, no. 12, pp. 13568–13613, Sep. 2022.
- [18] N. Krishnaveni, "Feature selection algorithms for data mining classification: A survey," *Indian J. Sci. Technol.*, vol. 12, no. 1, pp. 1–11, Jan. 2019.
- [19] A. Bommer, X. Sun, B. Bischl, J. Rahnenführer, and M. Lang, "Benchmark for filter methods for feature selection in high-dimensional classification data," *Comput. Statist. Data Anal.*, vol. 143, Mar. 2020, Art. no. 106839.
- [20] D. Watson, "Conceptual challenges for interpretable machine learning," *SSRN Electron. J.*, p. 65, 2022.
- [21] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong, "Interpretable machine learning: Fundamental principles and 10 grand challenges," *Statist. Surveys*, vol. 16, no. 1, pp. 1–85, Jan. 2022.
- [22] S. Chakraborty, R. Tomsett, R. Raghavendra, D. Harborne, M. Alzantot, F. Cerutti, M. Srivastava, A. Preece, S. Julier, R. M. Rao, T. D. Kelley, D. Braines, M. Sensoy, C. J. Willis, and P. Gurram, "Interpretability of deep learning models: A survey of results," in *Proc. IEEE SmartWorld, Ubiquitous Intell. Comput., Adv. Trusted Comput., Scalable Comput. Commun., Cloud Big Data Comput., Internet People Smart City Innov.*, Aug. 2017, pp. 1–6.
- [23] Y. Zhang, P. Tino, A. Leonardis, and K. Tang, "A survey on neural network interpretability," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 5, no. 5, pp. 726–742, Oct. 2021.
- [24] M. Pawlicki, A. Pawlicka, R. Kozik, and M. Choras, "Advanced insights through systematic analysis: Mapping future research directions and opportunities for xAI in deep learning and artificial intelligence used in cybersecurity," *Neurocomputing*, vol. 590, Jul. 2024, Art. no. 127759.
- [25] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *Proc. IEEE 5th Int. Conf. Data Sci. Adv. Analytics (DSAA)*, Oct. 2018, pp. 80–89.
- [26] U. Kamath and J. Liu, *Pre-Model Interpretability Explainability*. Cham, Switzerland: Springer, 2021, pp. 27–77.
- [27] C. E. Shannon, *The Mathematical Theory of Communication*. Champaign, IL, USA: University of Illinois Press, 1949.
- [28] R. A. Fisher, "Statistical methods for research workers," in *Breakthroughs in Statistics: Methodology and Distribution*, New York, NY, USA: Springer, 1970, pp. 66–70.
- [29] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, no. 2, pp. 179–188, Sep. 1936.
- [30] T. K. Ho, "Random decision forests," in *Proc. 3rd Int. Conf. Document Anal. Recognit.*, vol. 1, 1995, pp. 278–282.
- [31] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, Jun. 2001.
- [32] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Jul. 2011.
- [33] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach. Learn.*, vol. 63, no. 1, pp. 3–42, Apr. 2006.
- [34] W. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bull. Math. Biol.*, vol. 52, nos. 1–2, pp. 99–115, 1990.
- [35] M. Yeo, Y. Koo, Y. Yoon, T. Hwang, J. Ryu, J. Song, and C. Park, "Flow-based malware detection using convolutional neural network," in *Proc. Int. Conf. Inf. Netw. (ICOIN)*, Jan. 2018, pp. 910–913.
- [36] C. Goller and A. Kuchler, "Learning task-dependent distributed representations by backpropagation through structure," in *Proc. Int. Conf. Neural Netw. (ICNN96)*, vol. 1, 1996, pp. 347–352.
- [37] O. Maimon and L. Rokach, *Data Mining and Knowledge Discovery Handbook*, 2nd ed., New York, NY, USA: Springer, Jan. 2010.
- [38] I. N. da Silva, D. H. Spatti, R. A. Flauzino, L. H. B. Liboni, and S. F. dos Reis Alves, *Artificial Neural Networks: A Practical Course*. Cham, Switzerland: Springer, 2017.
- [39] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–20.
- [40] R. A. Fisher, "Iris," UCI Mach. Learn. Repository, 1988, doi: [10.24432/C56C76](https://doi.org/10.24432/C56C76).
- [41] S. Aeberhard and M. Forina, "Wine," UCI Mach. Learn. Repository, 1991, doi: [10.24432/C5PC7J](https://doi.org/10.24432/C5PC7J).
- [42] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Statist.*, vol. 32, no. 2, pp. 407–451, 2004. [Online]. Available: <http://www.jstor.org/stable/3448465>
- [43] "Dry bean," UCI Mach. Learn. Repository, 2020, doi: [10.24432/C50S4B](https://doi.org/10.24432/C50S4B).
- [44] "Rice (Cammeo and Osmancik)," UCI Mach. Learn. Repository, 2019, doi: [10.24432/C5MW4Z](https://doi.org/10.24432/C5MW4Z).
- [45] B. German, "Glass identification," UCI Mach. Learn. Repository, 1987, doi: [10.24432/C5WW2P](https://doi.org/10.24432/C5WW2P).
- [46] N. Moustafa, "A new distributed architecture for evaluating AI-based security systems at the edge: Network TON_IoT datasets," *Sustain. Cities Soc.*, vol. 72, Sep. 2021, Art. no. 102994.
- [47] M. Sarhan, S. Layeghy, N. Moustafa, and M. Portmann, "NetFlow datasets for machine learning-based network intrusion detection systems," in *Big Data Technology Application*. Cham, Switzerland: Springer, 2021.
- [48] M. Waskom, "Seaborn: Statistical data visualization," *J. Open Source Softw.*, vol. 6, no. 60, p. 3021, Apr. 2021, doi: [10.21105/JOSS.03021](https://doi.org/10.21105/JOSS.03021).



MAREK PAWLICKI received the Ph.D. (Eng.) degree. He holds an adjunct position with Bydgoszcz University of Science and Technology. He has been involved in a number of international projects related to cybersecurity, critical infrastructure protection, and software quality (e.g., H2020 SPARTA, H2020 SIMARGL, H2020 PREVISION, H2020 MAGNETO, H2020 Q-Rapids, and H2020 SocialTruth). He is the author of over 90 peer-reviewed scientific publications. His research interest includes the application of machine learning in several domains, including cybersecurity.

• • •