Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi
Pamukkale University Journal of Engineering Sciences

A comparative analysis on the reliability of interpretable machine learning

Yorumlanabilir makine öğrenmesinin güvenilirliği üzerine karşılaştırmalı bir analiz

Mustafa YILDIRIM[1], Feyza YILDIRIM OKAY[1], Suat ÖZDEMİR[2]*

[1]Department of Computer Engineering, Engineering Faculty, Gazi University, Ankara, Turkey.
mustafa.yildirim2@gazi.edu.tr, feyzaokay@gazi.edu.tr
[2]Department of Computer Engineering, Engineering Faculty, Hacettepe University, Ankara, Turkey.
ozdemir@cs.hacettepe.edu.tr

Received/Geliş Tarihi: 04.12.2022 Revision/Düzeltme Tarihi: 12.09.2023 doi: 10.5505/pajes.2023.49473
Accepted/Kabul Tarihi: 20.07.2023 Research Article/Araştırma Makalesi

Abstract

There is often a trade-off between accuracy and interpretability in Machine Learning (ML) models. As the model becomes more complex, generally the accuracy increases and the interpretability decreases. Interpretable Machine Learning (IML) methods have emerged to provide the interpretability of complex ML models while maintaining accuracy. Thus, accuracy remains constant while determining feature importance. In this study, we aim to compare agnostic IML methods including SHAP and ELI5 with the intrinsic IML methods and Feature Selection (FS) methods in terms of the similarity of attribute selection. Also, we compare agnostic IML models (SHAP, LIME, and ELI5) among each other in terms of similarity of local attribute selection. Experimental studies have been conducted on both general and private datasets to predict company default. According to the obtained results, this study confirms the reliability of agnostic IML methods by demonstrating similarities of up to 86% in the selection of attributes compared to intrinsic IML methods and FS methods. Additionally, certain agnostic IML methods can interpret models for local instances. The findings indicate that agnostic IML models can be applied in complex ML models to offer both global and local interpretability while maintaining high accuracy.

Keywords: Interpretable machine learning, default prediction, reliability, Jaccard index similarity, feature selection

Öz

Makine Öğrenmesi (ML) modellerinde genellikle doğruluk ve yorumlanabilirlik arasında bir denge vardır. Model daha karmaşık hale geldikçe, genellikle doğruluk artar ve yorumlanabilirlik azalır. Yorumlanabilir Makine Öğrenimi (IML) yöntemleri karmaşık ML modellerinin doğruluğunu korurken yorumlanabilirliğini sağlamak için ortaya çıkmıştır. Böylece, öznitelik önemi belirlenirken doğruluk sabit kalır. Bu çalışmada, SHAP ve ELI5 gibi agnostik IML yöntemleri ile içsel IML yöntemleri ve özellik seçimi (FS) yöntemlerinin öznitelik seçimi benzerliği açısından karşılaştırılmasını amaçlıyoruz. Ayrıca agnostik IML modellerini (SHAP, LIME ve ELI5) yerel öznitelik seçiminin benzerliği açısından kendi aralarında karşılaştırıyoruz. Şirket temerrüdünü tahmin etmek için hem genel hem de özel veri kümeleri üzerinde deneysel çalışmalar yapılmıştır. Elde edilen sonuçlara göre, bu çalışma öznitelik seçiminde içsel IML yöntemleri ve FS yöntemlerine kıyasla %86'ya kadar benzerlikler göstererek agnostik IML yöntemlerinin güvenilirliğini doğrulamaktadır. Ek olarak, bazı agnostik IML yöntemleri, modelleri yerel örnekler için de yorumlayabilmektedir. Sonuçlar, agnostik IML modellerinin, yüksek doğruluğu korurken genel ve yerel yorumlanabilirlik sağlamak için karmaşık ML modellerinde uygulanabileceğini göstermektedir.

Anahtar Kelimeler: Yorumlanabilir makine öğrenmesi, temerrüt tahmini, güvenilirlik, Jaccard dizin benzerliği, öznitelik seçimi

1 Introduction

The benefits of Machine Learning (ML) to human life, society, and the environment are indisputable. There are two main goals arising from the ML structure that are (i) ensuring high accuracy and (ii) providing interpretability by preventing the model from behaving like a black-box. However, there is a trade-off between the goals as seen in Figure 1 [1]–[3]. The higher the interpretability, the lower the accuracy, and vice versa. In other words, as ML methods become more complex, generally their accuracy increases and their interpretability decreases.

Most powerful but complex ML models such as Random Forests (RF), Deep Learning (DL), and Gradient Boosting Methods (GBM) generally have better performance at yielding highly accurate results on various real-world classification, regression, and prediction problems as compared to transparent ML models such as Linear Regression (LR), Decision Tree (DT), and Naive Bayes (NB). However, the behavior of complex ML models is often not transparent to their users. By acting like a black-box, they exclude users in the decision-making process. ML users are unaware of which particular decisions affect the results. In finance, for example, it is important for credit-based score models. Financial institutions must be fair as set forth by law as to whether to lend to individuals or companies. Based on this more detailed explanations of why borrowers' loans were declined - reason codes - are required.

Feature Selection (FS) becomes the focus of research areas of ML problems with the increase in the number of variables of datasets leading to high dimensional data. FS preprocesses the data to reduce feature size, which also allows for a certain level of interpretability. In addition, it also causes data loss, which may result in a decrease in the accuracy of the ML model.

Interpretable Machine Learning (IML) has emerged to mitigate problems arising from ML models acting like a

*Corresponding author/Yazışılan Yazar

black-box. It has the ability to explain or present the behavior of ML models. Thus, it provides a better understanding of how these models behave in predicting outputs [4]. In the literature, IML models can be classified as intrinsic IML models such as DT and Lasso and agnostic IML models such as SHAPley Additive exPlanations (SHAP) [5], ELI5 [6], and Local Interpretable Model-agnostic Explanations (LIME) [7].

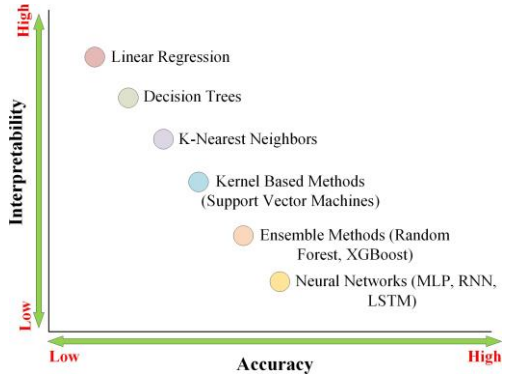

Figure 1. The trade-off between accuracy and interpretability [1]

In this paper, we focus on comparing agnostic IML methods with intrinsic IML methods and FS methods. All methods consider the effects of the features on the result. While FS methods commonly consider the effects of the features before training the model (pre-processing), IML methods consider it during (intrinsic IML) or after (agnostic IML) training the model. Here, we aim to validate the reliability of agnostic IML methods by comparing them with intrinsic and FS methods for similarities in attribute selection. Achieving high consistency also indicated that the reliability of agnostic IML methods is potentially increased. In addition, we compare agnostic IML models between each other for an instance company (for local interpretability). To the best of our knowledge, this is the first attempt to perform a comprehensive analysis of agnostic IML methods. Other contributions of the paper are listed below.

- Experiments are conducted in a detailed and comparative way in both public and private real-world datasets, providing a fine-grained examination of the behavior of the models.
- Experimental results of similarity of attribute selection confirm that agnostic IML methods are highly correlated with intrinsic IML and FS methods. Hence, IML methods can be applied to complex ML models to handle the accuracy-interpretability trade-off.
- Unlike FS and intrinsic IML methods, agnostic IML methods can potentially provide local interpretability as well as global interpretability. We analyze the local interpretability results by comparing agnostic IML methods among each other for both datasets.

The rest of the paper is organized as follows: Section II reviews the literature by classifying the existing methods according to their target domains and applications. Section III presents brief information about the presented approaches. Section IV explains the datasets, experimental setup, and experimental results for global and local attribute selections. Also, it provides a detailed discussion of the results. Lastly, Section V concludes the paper by giving some highlights about what we achieved.

2 Related Work

This study presents a comparative analysis of FS, intrinsic IML, and agnostic IML methods. All of them contribute to interpretability when applied to ML models in the pre-training, in-training, and post-training processes, respectively. Unlike existing studies, this study analyzes the similarity of attribute selection of these methods comprehensively. In the literature, FS and IML methods are examined separately. For this reason, we perform related works by grouping them according to the use of FS and IML methods.

2.1 Applications of Feature Selection

In the literature, FS methods have been generally used to improve the quality of attribute sets in different ML tasks and domains [8]. According to training data is labeled or not, FS methods can be categorized into three groups which are (i) supervised, (ii) unsupervised, and (iii) semi-supervised FS. In literature, there are comprehensive reviews that present detailed and comparative analysis of FS methods for supervised learning [9], [10], unsupervised learning [11], [12], and semi-supervised learning [13]. The existing FS methods for supervised learning can be further classified into filter methods, and wrapper methods.

By employing different techniques, numerous studies are presented in the literature to prove the benefits and success of FS methods on different real-world applications such as text mining [14], image processing [15], intrusion detection [16], information retrieval [17] by offering promising solutions on different financial [18], biological [19], medical [20], security [21], agricultural [22] and environmental [23] issues. Furthermore, more recent studies focus on combining FS with heuristic [24] and meta-heuristic [25] methods.

However, based on the commonly accepted consensus on FS methods, there is no so-called 'best method', thereby the researchers focus on seeking a good result by using new FS methods with different strategies [26], [27]: (i) combining more than one FS methods such as filter-filter or filter-wrapper, (ii) combining with other techniques such as feature extraction (iii) FS with an ensemble or heuristic methods, (iii) reinterpreting existing methods, (iv) adapting an existing method to a certain type of problems, (iv) creating a new method for unresolved problems.

As seen in Table I, FS methods are frequently used in finance as well as in other fields in the literature. These methods are especially used to improve financial predictions such as financial distress, credit risk evaluation, and financial crisis. Li et al. [28] aim to identify financial distress with SVM. Statistics-based wrapper FS is employed to determine the effective features. To that end, the authors first compare the statistics-based wrapper with filter FS methods and non-FS methods for SVM in financial distress prediction. Then, the proposed wrapper is conducted on some variants of SVM like linear SVM (LSVM), polynomial SVM (PSVM), Gaussian SVM (GSVM), and sigmoid SVM (SSVM). The experiments are applied to the data for financial distress prediction collected from Chinese public companies. Also, there are three benchmark FS methods including a wrapper, MDA, and Logit FS methods. Each FS is applied on each SVM variant, and the results are compared in terms of prediction performance and a two-tailed significance test. Cui et al. [29] introduce

Table I. FS Methods in Financial Problems

Reference	Problem	Datasets	FS methods	Benchmarks	ML models	Evaluation metrics
Li et al. [28]	Distress prediction	Data collected from Chinese public companies	Statistics based wrapper FS	A wrapper FS, MDA FS, Logit FS	LSVM, PSVM, GSVM, SSVM	Prediction performance, two-tailed significance test
Cui et al. [29]	Credit risk evaluation	P2P(DI), P2P200, P2P(R), CrowdFunding and BorrowerCredit	MSIEN	InElasticNet, ElasticNEt, Lasso, InLasso	SVM	Accuracy, convergence
Jadhav et al. [30]	Credit scoring	German, Australian and Taiwan credit datasets	Information gain directed FS	Baseline classifier, GA wrapper	SVM, KNN, NB	Accuracy, ROC curves
Liang et al. [31]	Credit scoring	German, Australian credit datasets and Chinese and Taiwanese bankruptcy datasets	GA and PSO wrapper based FS, t-Test, LDA and LR filter based FS	-	SVM, RBF kernel-SVM, NB, KNN, MLP Clas. and Reg. Tree	Accuracy, Type-I error
Sivasankar et al. [32]	Credit risk prediction	German, Australian and Japanese credit datasets	Ensemble with RS-FS	Ensemble without FS, ensemble with GR-FS	SVM, KNN, LoR, DT	Accuracy, AUC
Zhang et al. [33]	Stock prediction	Annual financial reports of A-Shares of the Shanghai Stock Exchanges	NoFS, CFS, PCA, CART, Lasso	-	LR, NB, BN, NN, SVM, J48, RF	Accuracy, number of features, precision
Lin et al. [34]	Bankruptcy prediction	Australian, German and Taiwan bankruptcy datasets	Information gain, GA-based FS	-	MLP, DT, SVM, KNN	Type-I error

multiple structural interacting elastic net (MSIEN) on SVM for financial credit risk evaluation. The experiments conducted on datasets of internet financing highlight the effectiveness of the proposed model according to other FS methods such as InElasticNet, ElasticNet, Lasso, and InLasso in terms of accuracy and convergence. Some studies focus on credit risk assessment by proposing information gain directed FS [30], two wrapper methods including GA and PSO and three filter methods t-test, LDA, LoR [31], and ensemble with a rough set based FS [32] on well-known classifiers such as SVM, KNN, NB, MLP by analyzing the evaluation metrics like accuracy, ROC curves, AUC values, and Type-I errors. Also, Zhang et al. [33] perform comprehensive and comparative analysis with different FS methods such as CFS, PCA, CART, and Lasso on various classifiers including LR, NB, BN, NN, SVM, J48, and RF to predict stock exchanges. Shanghai stock Exchange datasets are used for experiments. The results are evaluated in terms of accuracy, number of features, and precision. Lastly, Lin et al. [34] attempt to provide an effective bankruptcy prediction model while comparing GA method as a wrapper FS and information gain as a filter method. They apply the presented FS methods on existing classifiers consisting MLP, DT, SVM, KNN and evaluate them in terms of Type-I error criterion.

2.2 Applications of Interpretable Machine Learning

Although the term IML is relatively new, the problem of explaining expert systems dates back much further, mid-1970's. Afterward, for many years, the general trend has shifted to the development of new methods and algorithms with high predictive power [35], [36]. However, especially after 2016, interpretability has gained importance again with its widespread usage in critical areas such as health [37], autonomous systems [38], and finance [39].

Although intrinsic algorithms such as LR and DT provide interpretability inherently, they often suffer from poor performance than complex and hard-to-interpret models like deep neural network, random forest, and gradient boosting machine. For this reason, recent studies aim to ensure that powerful algorithms reduce their opaqueness while maintaining their prediction success.

Credit scoring has gained high popularity since the recent innovations in the field of artificial intelligence enable experts to make insightful and more accurate decisions. With IML methods, decisions potentially can be made both accurate and interpretable. However, there are limited studies addressing financial issues with IML in the literature. Table II shows the related works in the literature that use IML methods for financial prediction. Demajo et al. [39] propose a 360-degree explanation framework that enables three different explanations (global, local feature-based, and local instance-based) on XGBoost algorithm to predict credit scoring. It employs SHAP+GIRP for global explanations, Anchors for local feature-based explanations, and ProtoDash for local instance-based explanations. The model is applied on Home Equity Line of Credit (HELOC) and Lending Club (LC) datasets. The proposed model is compared with BRCG method, which generates Boolean rules to interpret the model globally. Also, three different evaluation approaches which are functionally-grounded, application-grounded, and human-grounded are adopted to analyze the proposed model in terms of consistency, simplicity, correctness, effectiveness, easy understanding, detail sufficiency, and trustworthiness.

TABLE II. IML Methods in Financial Predictions

Reference	Problem	Datasets	IML methods	Benchmark IML methods	ML models	Benchmark ML models	Explanation types	Evaluation metrics
Demajo et al. [39]	Credit scoring	HELOC, LC datasets	SHAP+GIRP, Anchors, ProtoDash	Boolean Rules via Column Generation (BRCG)	XGBoost	-	Local feature-based, Local instance-based, Global	Number of unique rules, Average number of rule conditions, Consistency of rules, Completeness rate
Bussman et al. [40]	Credit scoring	Data from ECAI	SHAPley values	-	XGBoost	Optimal LoR	Local	-
Ito et al. [41]	Text summarization	Posts on the Yahoo Finance Board, news articles	GINN	-	NN	Base MLP, plus MLP	Local	Correspondence
Cong et al. [42]	Portfolio management	CRSP, CRSP Compustat Merged, Financial Ratio Firm Level	AlphaPortfolio model with gradient-based methods, Lasso	-	LSTM, RNN, Transformer	-	Local	Out-of-sample Sharpe ratio, robustness
Grath et al. [43]	Credit risk assessment	HELOC dataset	Two weighted strategies for Counterfactual explanations	Baseline weighted Counterfactual explanations	LoR, GBM, SVC, MLP	-	Local	Predictive power, average size of generated counterfactual explanations
Ghosh et al. [44]	Financial stress	Financial stress variables regulated by the OFR	Permutation feature importance, LIME	-	EEMD-LSTM EEMD-Prophet	ARIMA, SARIMA, BSTS, MLP	Local, Global	Permutation importance, local level feature contribution
Babaei et al. [45]	Credit risk, expected return	financial indicators of a sample of 2049 Italian SMEs in 2018	SHAP	-	XGBoost	-	Local, Global	Feature importance
Tran et al. [46]	Financial distress prediction	Dataset includes companies in Vietnam from 2010 to 2021	SHAP	-	LR, SVM, DT, RF, XGBoost, ANN	-	Global	Feature importance
Ariza-Garzón et al. [47]	Credit risk scoring	LC dataset	SHAP	-	DT, RF, XGBoost,	LoR	Global	Feature importance, dependence, monotonicity
Misheva et al. [48]	Credit risk scoring	LC dataset	SHAP, LIME, Accumulated Local Effects (ALE) plots	-	LoR, SVM, RF, XGBoost, NN classifier	-	Local, Global	Feature importance, dependence,
Bracke et al. [49]	Default prediction	Snapshot of the residential mortgages in the UK	SHAP	-	Logit, GTB	-	Local, Global	Feature importance

Bussmann et al. [40] attempt to predict credit risk in peer-to-peer lending. To solve this problem, they propose XGBoost, which has high predictive success with limited interpretability. SHAPley value is employed to provide interpretability to the model in terms of the creditworthiness of each company. The proposed model is trained on data from European External Credit Assessment Institution (ECAI) and compared with optimal Logistic Regression (LoR) by evaluating their ROC curves. Ito et al. [41] propose gradient neural network (GINN) to visualize financial documents, thereby non-experts can easily understand the market sentiments. Two different datasets are obtained from the posts on the Yahoo Finance Board and Japanese financial news articles. Their proposal shows superiority when compared to the base MLP and plus MLP in terms of F scores. Also, human interpretability tests are conducted to measure the correspondence between sentiment scores obtained by GINN and perceptions of people. Cong et al. [42] introduce a reinforcement-based portfolio management model, called AlphaPortfolio, to increase the performance out-of-sample drastically and provide robustness. Then, the authors project the model onto a linear model by carrying out a polynomial sensitivity analysis that allows ML models to be more transparent and interpretable. The proposed model is conducted on three different WRDS databases which are CRSP, CRSP Compustat Merged, and Financial Ratio Firm Level. The experimental results show that the proposed model outperforms existing ML models especially when it is restricted by some financial constraints or restrictions.

Some IML methods like counterfactual explanation can be challenging when the number of features increased. Grath et al. [43] aims to overcome this problem by proposing positive counterfactual and two weighted counterfactuals which are feature importance and nearest neighbour based strategies. The experiments that are conducted on HELOC loan application datasets show that the weighted counterfactual generation strategy shows a better performance than the baseline counterfactual, by suggesting smaller counterfactuals while maintaining more interpretable decisions.

Ghosh et al. [44] emphasize the need for accurate prediction of financial stress and propose two granular hybrid predictive frameworks to discover the inherent pattern of financial stress across several critical variables and geography. The predictive structure utilizes the Ensemble Empirical Mode Decomposition (EEMD) for granular time series decomposition. Then, LSTM and Facebook's Prophet algorithms are invoked on top of the decomposed components to investigate the predictability of financial stress variables. Also, permutation feature importance and LIME methods are used to interpret the models and provide insights into the factors that contribute to financial stress.

Babaei et al. [45] propose an explainable AI model designed for analyzing SMEs and predicting their expected return based on credit risk and expected profitability. The proposed model employs SHAP enabling interpretable predictions from AI models both globally and locally. To validate the model, the authors extracted financial performance indicators from the annual balance sheets of 2049 SMEs. Tran et al. [46] compare the predictive performance of various machine learning algorithms including LoR, SVM, DT, RF, XGBoost, Artificial Neural Network (ANN) and use SHAP values to interpret the prediction results on a dataset of listed companies in Vietnam from 2010 to 2021. The experimental results reveal that XGBoost and RF models outperform other algorithms.

Ariza-Garzón et al. [47] conduct a comparative analysis of machine learning algorithms, including DT, RF, and XGBoost, against LoR models for predicting personal loans from the LC company. They further assess the contribution of variables in the models using SHAP and LIME methods. The outcomes demonstrate that the application of SHAP to machine learning methods significantly enhances the interpretability of these models, capturing nonlinear relationships more effectively compared to the traditional LoR model. Similarly, Misheva et al. [48] present a similar study on the LC dataset. The authors show that the explanatory results obtained are robust and coherent with logical financial explanations. In this study, the SHAP and LIME methods are employed to interpret the results of the machine learning models.

Bracke et al. [49] employ the Linear Logistic Regression (Logit) and GBM to predict mortgage loan defaults in the UK. The authors introduce a novel approach called quantitative input influence (QII), which evaluates the contribution of input variables to the target variable by computing Shapley values. Through this method, the authors demonstrate the ability to offer a comprehensive explanation of the variable's impact on various customer groups, providing detailed insights into the degree of influence for each group.

3 Background Information

In this section, we give additional information about FS and IML approaches by examining and categorizing them into sub-methods.

3.1 Feature Selection

Over the past few years, the dimensionality of data has been growing exponentially, causing serious problems to existing learning methods like curse of dimensionality. To address this problem, FS methods have become popular by reducing dimensionality for better performance, lower computational cost, and better interpretability. The aim of FS is to select a subset of relevant features that represent the data in the best way, thereby diminishing the effects of irrelevant or redundant data and constructing simpler and comprehensible models [50], [51].

FS process consists of four main steps. These steps are subset generation, subset evaluation, stopping criterion, and result validation. In the FS process, first, a subset of features is generated from the original dataset. The generation process depends on the state space search strategy. After the selection of the candidate subset, it is evaluated using certain methods. Subset generation and evaluation steps repeat until the stopping criteria are met. Hence, the best candidate selected features are determined. Lastly, the subset with these features is validated on an independent dataset [8].

In this study, we only focus on supervised FS methods, since our problem is a supervised learning problem. Depending on the evaluation criteria, these methods can be classified into two groups [9], [10], [52]: (i) filter methods based on statistical information, (ii) the wrapper methods that try to

achieve the best prediction performance by evaluating by using a predetermined ML algorithm.

3.1.1 Filter Methods

In filter methods, which are based on statistical information, the selection process is only performed based on statistical measurements like distance, frequency, or consistency without using any ML algorithm. These methods first obtain a score according to the evaluation criteria for each feature. Then, they form a subset of the features among the scores with the highest value. In these methods, the size of the attribute subset is determined according to the minimum sub- score value. Although there are many different filter methods, Mutual Information (MI), Chi-square (X2), F-score, T-score, Information Gain (IG), and relief are among the well-known methods in the literature [52].

MI is a filter-based method based on probability theory. It is a measure of mutual information between two variables. In other words, it is a measure of the amount of information about a random variable by observing another random variable. It uses entropy. Entropy is a measure of uncertainty and is calculated as in Equation (1):

$$Entropy(s) = \sum_{i=1}^{c} -p_i log_2 \, p_i \tag{1}$$

where s denotes the training dataset, and c is the number of different values of the target. p_i is the probability of the target variable in training data of class i. MI tends to select features with high entropy.

3.1.2 Wrapper Methods

In wrapping methods, after the generation of a subset of features, ML algorithms are employed instead of statistical criteria such as distance, frequency, and consistency. The subset of features that makes the best predictive performance of ML is selected. Since wrapper methods consider the performance of the selected features unlike filter methods, wrapper methods achieve better performance than filter methods in exchange for computation cost. The most commonly used wrapper methods in the literature are Sequential Forward Selection (SFS), Sequential Backward Selection (SBS), Sequential Forward Floating Selection (SFFS), and Sequential Backward Floating Selection (SBFS) [9].

SFS starts with an empty subset of attributes and decides whether to add a new attribute at each step. The attribute is added to the subset if it enhances the performance of the model. It is extracted if it lowers success. Until every attribute has been tested, this procedure continues.

3.2 Interpretable Machine Learning

IML is a more recent and general concept than FS. In some studies, it covers FS as a technique where interpretability is achieved before building the model [53]. Note that, dimensionality reduction methods like FS or feature extraction can provide interpretability since the outcomes are intuitively explained by selected attributes [54].

As seen in Figure 2, IML enables users and part of internal systems to be more transparent and allows them to explain how they make decisions [55]. This concept also overlaps with the explainable artificial intelligence concept, which is frequently used in the literature. While both concepts are similar and serve the same purpose, there is little difference between the meanings. Interpretable systems become explainable when their inner operations are intelligible, in other words, understandable by a human. As in the ML community, the term 'interpretable' is more commonly used [35], so we prefer to use the concept of IML. But still, we utilize both concepts interchangeably in this study.

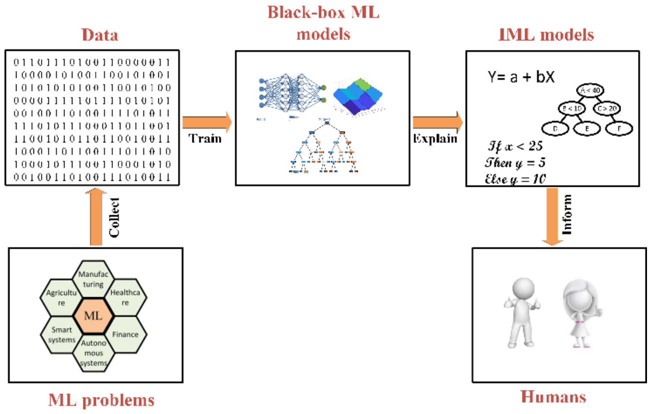

Figure 2. The general view of IML [55]

According to the study of [56], the goals of interpretability are to achieve (i) accuracy that represents the similarity between explanations and model predictions, (ii) readability that represents the simplicity of understanding the behavior of the model and (iii) efficiency that represents the time constraint to comprehend evaluations.

In addition, the quality of interpretability can be assessed by some properties described in the study of [57].

- Accuracy: It refers to the generalization of explanations of existing decisions to cover unseen instances.
- Fidelity: It describes the degree of how well an explanation expresses the behavior of ML model.
- Consistency: It measures the degree of difference between explanations in case different models are trained to fulfill the same task.
- Stability: It is a degree of difference between an application for similar instances. Apart from consistency, this property tackles with explanations obtained by the same model.
- Comprehensibility: It is related to the readability and size of explanations. In other words, it tries to measure how well explanations are understood by a human.
- Certainty: It assesses the reflection of explanations about the certainty of the ML model.
- Degree of Importance: It measures the degree of how well explanations present the importance of features.
- Novelty: It considers whether an explained instance is included in the training set, thereby evaluating the certainty.
- Representativeness: It assesses the representativeness of a model by considering whether it covers the behavior of the whole or part of the model.

According to when the technique is applicable, the IML methods in the literature can be classified into intrinsic and agnostic IML methods [53]. Intrinsic models inherently restrict the complexity of ML models during the period of training. Due to the restriction during training, attributes are intuitively listed by feature importance, which may provide also interpretability. On the other hand, agnostic models provide analysis and interpretation of the model with certain methods after the

training process.

Also, according to the scope of interpretability methods can be classified as global and local interpretability. Local interpretability provides a local understanding of why and how specific predictions can be made depending on an instance [39]. Local instance-based explanations focus on providing interpretability by looking at particular single or multiple instances.

SHAP [5], counterfactual explanations [58], and LIME are some popular methods presented for individual predictions. On the other hand, global interpretability aims to describe the model as a whole. More specifically, it requires comprehending decisions, features, structures, and each learned component such as weights and parameters. Especially, it seeks to answer the questions of which features are more important and what kind of interactions are realized among them.

3.2.1 SHAP

It is an agnostic IML algorithm proposed by Lundberg and Lee [5] in 2017. It is developed based on the SHAPley value which is one of the popular game theory methods. The purpose of SHAP is to describe how each attribute of a sample affects the prediction. The prediction is based on the SHAPley value that was developed to measure the marginal contribution of each player to the score in a team game. This is accomplished by calculating the difference in the score resulting from replacing each team member with a random player one at a time. In this way, the marginal contribution of each player to the score can be measured. Since each player is eliminated one by one, SHAPley value is calculated exponentially as the number of players on the team increases. The SHAP model treats each player as an attribute in a similar manner. It aims to provide the explainability of the method by quantifying the contribution of each attribute to ML algorithm prediction.

SHAP integrates LIME and SHAPley values and can be formalized as follows [5]:

$$g(z') = \Phi_0 + \sum_{j=1}^{M} \Phi_j z'_j \qquad (2)$$

where g is an explanatory model, M is the maximum number of attributes, and $z' \in [0,1]^{M}$ is the simplified new dataset. Lastly, $\Phi_j \in R$ is the contribution of the j attribute, which means that it is the SHAPley value. Here, SHAPley value, Φ_j can be calculated as in Equation (3) [5]:

$$\Phi_j = \sum_{S \sqsubseteq F \setminus \{j\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} [f(S \cup \{j\}) - f(S)] \qquad (3)$$

where Φ_j denotes the difference between the prediction with attribute j and the prediction without attribute j. Thus, the marginal contribution of the j attribute is determined. $f(S \cup \{j\})$ is the new set after including attribute j. F shows all theattributes. Lastly, $S \sqsubseteq F \setminus j$ shows possible subsets excluding the attribute j.

3.2.2 LIME

LIME [7] method is used to interpret the single instances of ML algorithms in a data set. It explains the single instances by using an interpretable surrogate model. This surrogate model is an interpretable model such as LR and DT and employs a heuristic approach. It considers ML algorithms as a black box and analyzes inputs and predictions of ML algorithms. Its aim is to understand why ML makes a certain prediction for a sample. To do so, it generates new samples that are similar to the chosen ones. They are weighted according to the degree of closeness. LIME is trained with the surrogate model with the new weighted data set. Interpretability is ensured by the weights given to the variables by the surrogate model.

LIME is mathematically formulated as shown in Equation (4):

$$\exp(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega_g \qquad (4)$$

where x is the sample to be explained. g is the surrogate model such as LR or DT, f is the actual model that makes the predictions such as RF, and π_x represents the closeness measure of the sample x. L is the function that minimizes the loss function, such as the least mean square error. Ω_g is used to keep the number of attributes of the surrogate model low. Note that, the number of attributes is usually determined by the user.

3.2.3 ELI5

ELI5 [6] uses the permutation importance or mean decrease accuracy method to provide interpretability to complex algorithms. This method eliminates each attribute from the dataset, retrains the model, and weights the attributes according to the decrease in the success measure used by the model, such as F1 score, R-squared or accuracy. Since it is repeated for each attribute, it is a computationally expensive algorithm. However, it is widely used because it supports Python libraries such as Sklearn and Keras.

4 Experimental Analysis and Discussion

In this study, we perform a comparative analysis by grouping the methods into three categories as FS, intrinsic IML, and agnostic IML methods. We intend to measure the similarities in the selection of attributes. Our motivation is to increase the reliability of agnostic IML methods by confirming the similarity of attribute selection with FS and intrinsic IML methods. Thus, agnostic IML methods will be able to provide interpretability by applying them to highly accurate complex models.

The experimental study is separated into two sub-studies comparing global and local attribute selections for two different datasets. The pseudocode of the study is given in Table III. In the aim of measuring the similarity among different models, the first study includes the comparison of the global attribute selection results for FS, intrinsic IML, and agnostic IML models. In the second study, the similarity of local attribute selection is measured and compared with three different IML methods.

The pseudocode outlines the steps of the study. In summary, the pseudocode begins by reading two different datasets and imputing missing data with '0'. The dependent variable for predicting default is separated from the independent variables. For each dataset, the threshold values of 25% and 50% are determined and listed. The study encompasses both global and local similarity calculations, which are coded separately.

Table III. The pseudocode of the study

```
Read the datasets (Turkish and Polish datasets).
Perform imputation on missing data by replacing them with 0.
Separate the independent variable.
For each dataset:
    Determine the number of selected attributes (either 25% or 50%).
#Global similarity
For each dataset:
    For each threshold value:
        Run MI algorithm.
        Run SFS algorithm.
        Run Lasso algorithm.
        Run DT algorithm.
        Create prediction model using RF for SHAP and ELI5:
            Run SHAP algorithm.
            Run ELI5 algorithm.
        Create a dataframe of selected attribute by the 6 algorithms.
        Generate a table of the most commonly selected attributes by the 6 algorithms.
        Compare the attribute dataframe using Jaccard similarity.
        Visualize the comparison results using a heatmap.
#Local similarity
For each dataset:
    Select randomly 10 default companies and 10 non-default companies.
    For each threshold value:
        Create a prediction model using RF:
            For each selected company:
                Run SHAP algorithm.
                Run LIME algorithm.
                Run ELI5 algorithm.
        Create a dataframe of selected attributes by the 3 algorithms.
        Generate a table of the most commonly selected attributes by the 3 algorithms based on the threshold value.
        Compare the attribute dataframe using Jaccard similarity.
        Calculate the average of the comparison results and visualize them using a heatmap.
```

To measure global similarity, nested loops are created for each dataset and threshold value. For each dataset and threshold value, the MI, SFS, Lasso, DT, SHAP, and ELI5 algorithms are executed in sequence, through the entire dataset. Since SHAP and ELI5 algorithms require a trained ML model, the RF algorithm is first trained on the entire dataset, and then SHAP and ELI5 are executed. The selected attributes by the six algorithms are stored in a dataframe, and tables of the most commonly selected attributes by the six algorithms are generated. The attribute dataframes are compared using Jaccard similarity, and the results are visualized using a heatmap. This process is repeated for each dataset and threshold value.

For local similarity, intrinsic models are not applicable, thus 20 companies (10 default and 10 non-default) are randomly selected. For each dataset, 20 companies are selected. For each threshold value, the RF algorithm is executed. For each dataset, threshold value, and selected company, SHAP, LIME, and ELI5 algorithms are executed. The selected attributes by the three algorithms are stored in a dataframe. The attribute dataframes are compared using Jaccard similarity. The average of the comparison results for the 20 companies is calculated and visualized using a heatmap.

4.1 Dataset

In this study, the company default prediction problem is addressed. By using the variables from balance sheet data, company default is predicted. Here, company default is the inability of the company to pay its debt on time.

Experimental studies are conducted on two different datasets which are Polish and Turkish datasets. Both datasets consist of values of ratios calculated from company balance sheets such as net profit / total assets, total assets / total assets, working capital / total assets, and current assets / short-term assets. Also, all values of each attribute are continuous. Therefore, the attributes of these datasets are homogeneous. Our first dataset [59] is public and belongs to Polish companies operating between 2000-2012. It has 64 attributes and 43405 instances. Our second dataset is private and randomly selected from Turkish companies operating between 2015-2017. It has 74 attributes and 43318 instances. Since the missing value ratio is below 1% in both datasets and all attributes are numeric, we resolved the missing data problem by assigning a value of '0'.

TABLE IV. Default rates for Polish and Turkish companies

	Number of attributes	**Number of instance**	**Default rate**	**Data type**
Polish dataset	64	43405	0.0481	Numeric
Turkish dataset	74	43318	0.0719	Numeric

The default rates for the datasets are given in Table IV. As can be seen from the default rates, the datasets are unbalanced.

4.2 Experimental Setup

All experiments have been performed on a notebook with Intel Core i7-7600U CPU 2.9 GHz processor and 15.9 GM RAM. The models have been developed on Python 3.7.6

version. Also, Python's Sklearn, SHAP, LIME, and ELI5 libraries are employed in the experiments.

The FS and intrinsic IML methods are generally used with default values from the Python Sklearn library. The number of neighbors for MI is 3. K Neighbors Classifier is used for SFS. The number of neighbors for SFS is set to 3 as well. Accuracy is determined for scoring. For Lasso, L1 was chosen as the penalty value of the LoR classifier. The regularization strength value is set to 1. Liblinear is used as the solver. In DT classifier, Gini is used for criterion, min samples split is set to 2, and min samples leaf is set to 1.

4.3 Evaluation Criteria

There are various similarity measurement methods used for different purposes in the literature. Jaccard index similarity [60] is determined as the comparison metric, since the lists of attributes selected for the prediction are lists of the same length containing 0 and 1. As shown in Equation (5), Jaccard index is calculated for infinite sets by dividing the intersection of the sets by the union of the sets.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \quad (5)$$

where $J(A,B)$ represents Jaccard index and takes values between 0 and 1 inclusive. A and B represent two finite sets. The closer the Jaccard index is to 1, the greater the similarity.

Cosine similarity [61] is another commonly used similarity algorithm in the literature, which can be employed to enhance the reliability of the IML results. It is a metric used to measure similarity between vectors, generating a similarity score based on the angle between vectors. However, in this study, we solely conduct similarity analyses using Jaccard similarity to maintain the comprehensibility of the paper and avoid overwhelming the readers with excessive similarity results.

4.4 Comparison of Similarity of Global Attribute Selection

In the experiments for global interpretability, we categorize the methods into three groups: FS, intrinsic IML, and agnostic IML methods. In each category, two algorithms are selected. Accordingly, MI and SFS are chosen as FS methods, DT, and Lasso are selected as intrinsic IML models, and SHAP and ELI5 are determined as agnostic IML models. The goal is to demonstrate the consistency of attribute selection between agnostic IML methods and the FS and intrinsic IML methods. We compare the globally selected attributes by each algorithm in terms of attribute selection. This comparison is performed separately for both datasets.

Agnostic IML models are employed to determine the feature importance based on the ML algorithm results. Specifically, attribute selection is carried out using a threshold value obtained from the feature importance determined by the agnostic models. For this purpose, we utilize the RF algorithm as an ML algorithm since it is known for its high accuracy and complex interpretability. Hence, it serves as a suitable candidate for agnostic models to provide interpretability while maintaining high accuracy.

The results are obtained using two different threshold values, set at 25% and 50% of the total number of attributes. This corresponds to 16 and 32 attributes for the Polish dataset, and 19 and 37 attributes for the Turkish dataset, respectively. Subsequently, the similarity of the globally selected attributes is measured and compared among the previously determined algorithms.

4.4.1 Experimental Results for Global Interpretability

The selected six algorithms are run twice for both datasets. Table V shows the global attribute selection with a 25% threshold for each algorithm. Additionally, Figures 3 and 4 depict the similarity matrices for global attribute selection with thresholds of 25% and 50%, respectively. These matrices illustrate the similarity ratios calculated based on the Jaccard index of the attributes selected by the algorithms.

Statistical significance tests are performed for all experimental analyses conducted using the Jaccard index. The R language is utilized for statistical significance testing, and since the number of variables is small, the 'exact' method is applied. The statistical significance of the similarity of all models is assessed, and the p-values of the statistical significance tests for all models are found to be less than 0.05, indicating that the similarities are statistically significant.

4.4.2 Discussion for Global Interpretability

Table V shows the attributes selected by each algorithm with 25% threshold value. The attributes are listed so that the most selected attributes are located above in the table. As seen in this table, many attributes are selected by more than one algorithm. For the Polish dataset, 25 out of 64 attributes are chosen by at least two algorithms, while 24 attributes are not selected by any algorithm. Similar results are obtained within the Turkish dataset. Additionally, it is noteworthy that the relatively new SHAP and ELI5 algorithms show similar attribute selections compared to FS and intrinsic methods. These findings serve as strong indicators that enhance the reliability of agnostic IML models.

The similarities of the selected attributes are compared using the Jaccard index. In Figure 3 and Figure 4, Jaccard similarity matrices are calculated for both datasets with 25% and 50% threshold values, respectively. The highest similarity of 0.86 is measured between the DT and ELI5 algorithms with a 25% threshold value for the Polish dataset.

When comparing the Turkish dataset with the Polish dataset, fewer similarities are observed in the Turkish dataset due to its larger number of attributes. Having more attributes increases the likelihood of correlations between them, resulting in lower similarity. This situation may occur because algorithms randomly select among the attributes with high correlation during attribute selection.

Comparing the 25% and 50% threshold values, the 25% attribute selection shows higher similarity. The main reason for this is that the selected attributes (25%) and non-selected attributes (75%) cause an imbalance in favor of the non-selected ones. This imbalance leads to an increase in similarity. When the threshold is set to 50%, this imbalance

Table V. Global attributes with 25% of threshold.

Polish dataset							Turkish dataset						
	FS		IML					FS		IML			
			Intrinsic		Agnostic					Intrinsic		Agnostic	
Attributes	MI	SFS	DT	Lasso	SHAP	ELI5	Attributes	MI	SFS	DT	Lasso	SHAP	ELI5
Attr46	1	1	1	1	1	1	**L13**	1	0	1	1	1	1
Attr22	1	1	1	1	1	1	**F3**	1	0	1	0	1	1
Attr24	1	1	1	0	1	1	**L3**	1	0	1	0	1	1
Attr42	1	0	1	1	1	1	**T4**	1	0	1	0	1	1
Attr39	1	1	1	0	1	1	**F27**	1	0	1	0	1	1
Attr27	1	0	1	0	1	1	**F1**	1	1	0	0	1	1
Attr58	0	1	1	0	1	1	**F14**	0	0	1	1	1	0
Attr26	1	0	0	1	1	1	**F21**	1	0	1	0	1	0
Attr35	1	0	0	1	1	1	**F13**	0	1	0	1	1	0
Attr40	0	0	1	1	0	1	**F26**	1	0	0	0	1	1
Attr16	1	0	0	1	1	0	**P3**	1	1	0	0	1	0
Attr41	1	0	1	0	1	0	**F2**	1	0	0	0	1	1
Attr29	0	0	1	1	0	1	**L12**	1	0	0	0	1	1
Attr13	1	1	0	0	1	0	**L10**	1	0	1	0	0	1
Attr38	0	0	0	1	1	0	**L9**	0	1	1	0	0	1
Attr48	0	0	0	1	1	0	**P17**	0	0	1	0	1	1
Attr51	0	1	0	1	0	0	**F19**	0	0	1	0	1	1
Attr56	0	0	1	0	0	1	**F18**	0	1	0	1	0	0
Attr1	0	1	0	1	0	0	**F28**	0	1	0	1	0	0
Attr34	0	0	1	0	0	1	**P4**	1	1	0	0	0	0
Attr11	0	0	0	1	1	0	**F25**	0	1	0	1	0	0
Attr21	0	0	1	0	1	0	**P18**	0	1	1	0	0	0
Attr19	1	1	0	0	0	0	**P6**	0	1	0	0	1	0
Attr6	0	1	0	0	0	1	**L1**	1	1	0	0	0	0
Attr9	0	0	0	1	0	1	**P24**	0	1	0	1	0	0
Attr3	0	0	0	1	0	0	**T3**	0	0	1	0	0	1
Attr59	0	1	0	0	0	0	**F15**	0	1	0	1	0	0
Attr4	0	1	0	0	0	0	**P12**	0	1	0	0	1	0
Attr5	0	0	1	0	0	0	**T5**	0	0	1	0	0	1
Attr55	0	0	0	0	0	1	**F6**	0	1	0	1	0	0
Attr50	0	1	0	0	0	0	**T7**	0	0	1	0	0	1
Attr45	1	0	0	0	0	0	**T8**	0	0	1	0	0	1
Attr25	1	0	0	0	0	0	**L5**	0	0	0	1	1	0
Attr44	0	0	1	0	0	0	**T9**	0	1	0	1	0	0
Attr15	1	0	0	0	0	0	**P13**	1	0	0	0	1	0
Attr20	0	1	0	0	0	0	**P23**	0	0	1	0	0	0
Attr23	1	0	0	0	0	0	**T6**	0	0	0	0	0	1
Attr37	0	1	0	0	0	0	**P15**	0	0	1	0	0	0
Attr36	0	0	0	1	0	0	**P14**	1	0	0	0	0	0
Attr60	0	1	0	0	0	0	**F24**	0	1	0	0	0	0
							P11	0	0	1	0	0	0
							P8	0	0	0	1	0	0
							P7	0	1	0	0	0	0
							P5	1	0	0	0	0	0
							L2	1	0	0	0	0	0
							F20	0	0	0	1	0	0
							F17	0	0	0	1	0	0
							F16	0	0	0	1	0	0
							F12	0	0	0	1	0	0
							F9	1	0	0	0	0	0
							F5	0	0	0	1	0	0
							F4	0	0	0	1	0	0
							L8	0	0	0	1	0	0
							L4	0	1	0	0	0	0
							T2	0	0	0	0	0	1

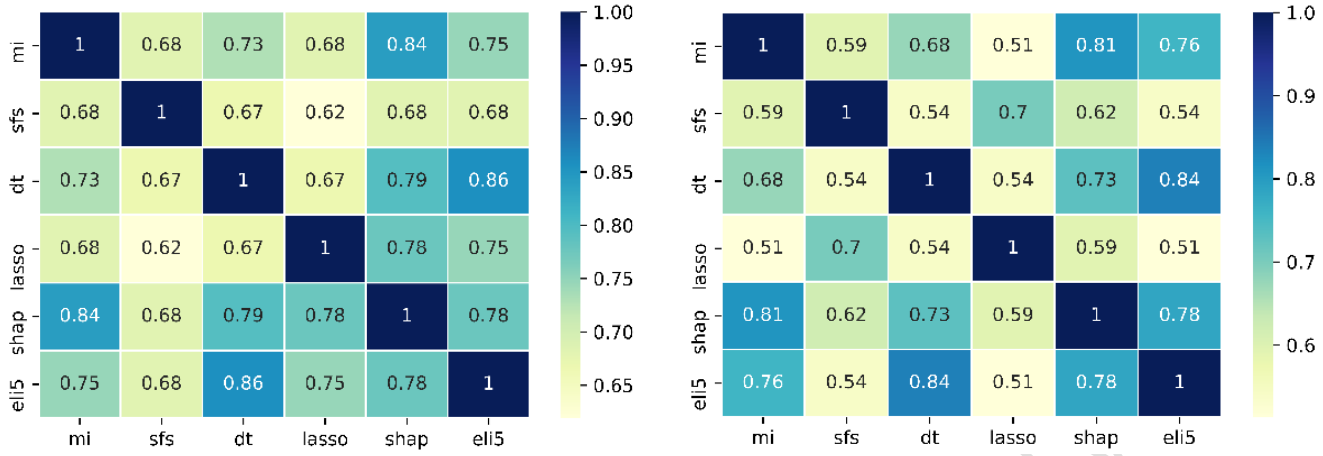

Figure 3. Jaccard index similarity of global attribute selection with 25% of threshold for a) Polish dataset b) Turkish dataset

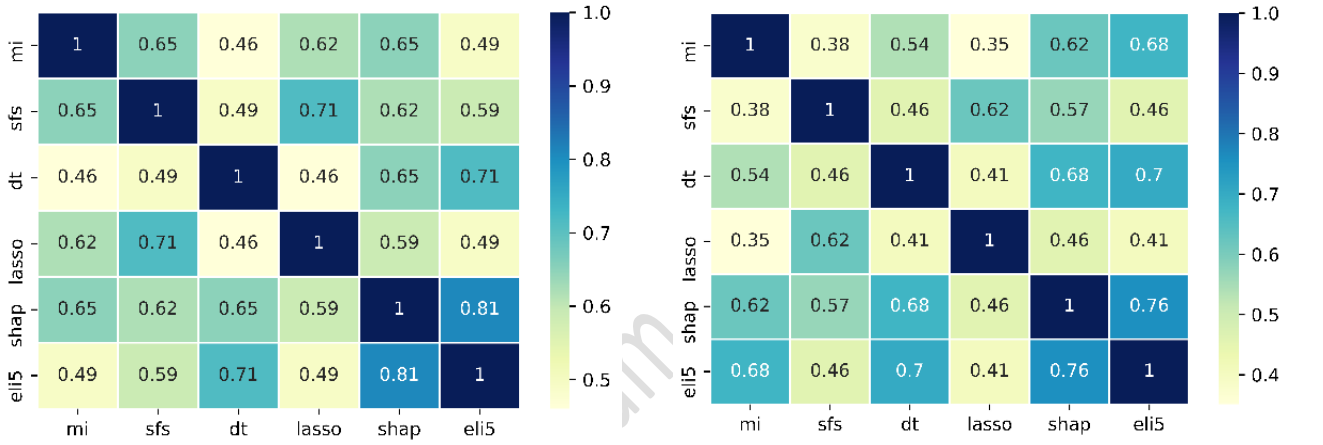

Figure 4. Jaccard index similarity of global attribute selection with 50% of threshold for a) Polish dataset b) Turkish dataset

disappears, and the attribute selection similarity results at 50% still indicate that IML algorithms are sufficiently similar to other algorithms.

Another important finding is that the similarity of SHAP and ELI5 algorithms becomes high for all experimental studies. This demonstrates that both agnostic models consistently produce similar results.

As seen in the matrices, there is a high similarity between agnostic IML models and their counterparts. This enables us to address the opacity issue of complex models, which are often preferred for prediction tasks due to their highly accurate prediction capability. The results of this study demonstrate that interpretability in complex models can be made reliable with agnostic models such as SHAP and ELI5.

As explained in Section II, complex ML algorithms are applied to various problems. However, their major drawback is poor interpretability. In recent years, IML algorithms have been proposed to maintain accuracy while prioritizing interpretability, especially in domains where interpretability is crucial, such as healthcare, automation, and finance.

The experimental results demonstrate a high level of consistency, which enhances the reliability of agnostic IML methods. Considering the expected increase in the use of IML models in the future, we believe that this study will make a valuable contribution to the literature.

4.5 Comparison of Similarity of Local Attribute Selection

Agnostic IML models can provide both local and global interpretability, unlike intrinsic models, which can only offer global interpretability. Local interpretability refers to explaining the behavior of a single instance rather than representing the entire dataset. This enables us to interpret the results for each instance within the dataset. For example, we can explain the default prediction for a single company using our dataset. Consequently, in the second part of the study, we aim to measure the similarity of agnostic models for local interpretability.

For the experiments on local interpretability, we consider the LIME algorithm in addition to the SHAP and ELI5 algorithms. We use the RF as our ML model. The experiments are conducted on 10 default and 10 non-default companies, randomly selected from both datasets. These companies are then evaluated for their local interpretability using the three agnostic models.

During the similarity analysis, the threshold value is set to 25%. This means that 25% of attributes are selected based on their feature importance to perform the local interpretability analysis.

4.5.1 Experimental Results for Local Interpretability

Table VI presents the local attribute selection with a 25% threshold value for a specific company in both datasets. The table shows the positive and negative effects of the attributes on the result, along with the coefficients of these effects for each algorithm. For this sample company, the effect direction and impact weight of the attributes selected by each agnostic IML model are depicted in Figures 5 and 6, respectively.

To assess the similarity of the local attribute selection for 20 different instance companies, the Jaccard index similarity is separately measured for the three agnostic models. Subsequently, the average Jaccard index similarity ratios for the selected companies from each dataset are calculated. In Figures 5 and 6, the average similarity ratios of these algorithms are presented.

4.5.2 Discussion for Local Interpretability

According to Table VI, it is seen that the selected and non-selected attributes by the three algorithms are similar. A comparison between the results in Table V and Table VI, where the global and local interpretability results are listed, shows that the attributes selected locally differ from the ones selected globally. This indicates that interpretability is tailored to capture different details locally, which is one of the major contributions of agnostic IML models.

The matrices in Figures 5 and 6 illustrate the results of local similarity of IML agnostic models. As FS and intrinsic IML models generally do not provide local interpretability, comparisons between these models could not be made. However, when all matrices are examined collectively, it becomes apparent that IML methods exhibit sufficient similarities.

TABLE VI. Local attributes with 25% of threshold for an instance company

Polish dataset				Turkish dataset			
Attributes	**SHAP**	**LIME**	**ELI5**	**Attributes**	**SHAP**	**LIME**	**ELI5**
Attr27	1	1	1	F27	1	1	1
Attr18	1	1	1	P6	1	1	1
Attr46	1	1	1	L12	1	1	1
Attr38	1	1	1	F2	1	1	1
Attr10	1	1	0	L3	1	0	1
Attr53	1	0	1	F4	0	1	1
Attr49	1	1	0	P3	1	0	1
Attr35	1	0	1	P13	1	0	1
Attr20	1	0	1	P17	1	0	1
Attr21	1	0	1	T4	1	0	1
Attr26	1	0	1	F3	1	0	1
Attr48	1	1	0	F1	1	0	1
Attr51	0	1	0	L13	1	0	1
Attr56	0	1	0	L10	1	0	1
Attr58	0	0	1	F26	0	1	1
Attr43	0	0	1	F25	0	1	0
Attr42	0	1	0	P24	1	0	0
Attr39	1	0	0	P23	0	1	0
Attr1	0	1	0	P12	1	0	0
Attr34	0	0	1	P4	0	1	0
Attr31	0	1	0	T5	0	0	1
Attr3	0	1	0	T8	1	0	0
Attr25	0	1	0	F21	1	0	0
Attr24	0	0	1	F24	0	0	1
Attr22	0	0	1	F23	0	1	0
Attr19	0	1	0	F22	0	1	0
Attr16	1	0	0	L4	0	1	0
Attr15	0	0	1	F20	1	0	0
Attr13	1	0	0	F18	0	1	0
Attr9	1	0	0	F17	0	1	0
Attr6	0	1	0	F15	0	1	0
Attr59	0	0	1	F14	0	1	0
				F13	0	1	0
				F12	1	0	0
				F6	0	1	0
				L8	0	1	0
				L6	0	0	1
				T1	0	0	1

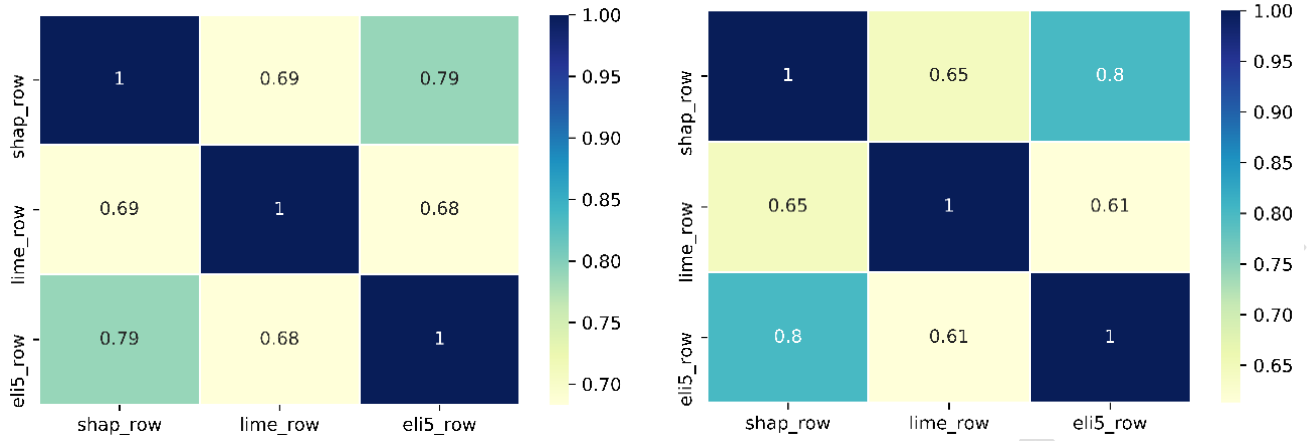

Figure 5. Jaccard index similarity of local attribute selection with 25% of threshold for a) Polish dataset b) Turkish dataset

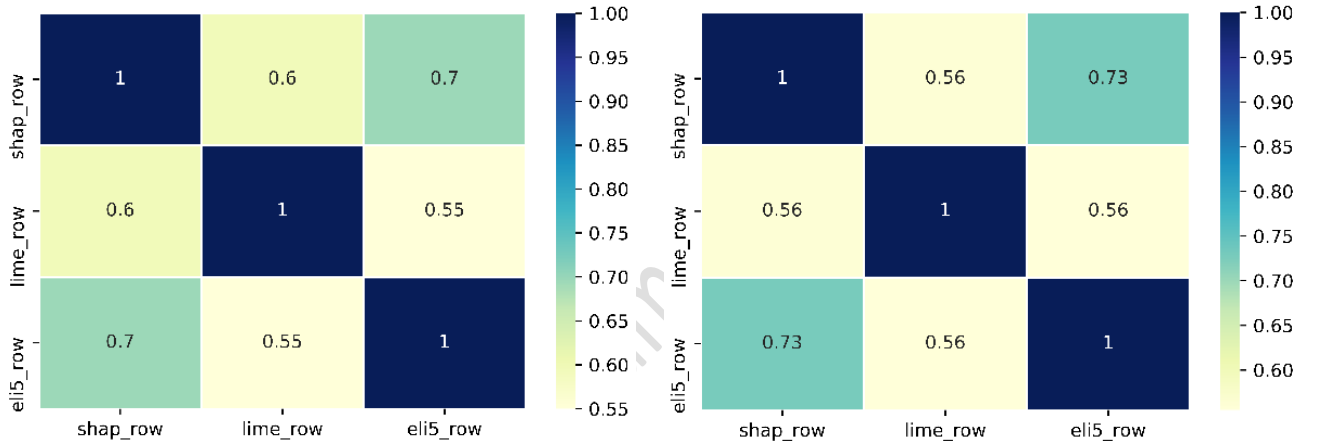

Figure 6. Jaccard index similarity of local attribute selection with 50% of threshold for a) Polish dataset b) Turkish dataset

5 Conclusion

ML algorithms have gained significant attention in different prediction problems recently. However, achieving high accuracy is not enough for ML algorithms to make insightful decisions for especially critical areas like finance or healthcare. Arising criticism about these algorithms is about their black- box wise behavior in the decision-making process. In recent years, IML algorithms have emerged to solve this problem by offering both high accuracy and interpretability when applied to complex ML methods. In this study, we aim to measure the reliability of agnostic IML algorithms by comparing them with FS and intrinsic IML methods. Results are evaluated by Jaccard index similarity. As it is expected, results clearly show that the agnostic IML methods produce similar results to FS methods and intrinsic IML methods. In other words, especially agnostic IML techniques can potentially provide interpretability as well as high accuracy for complex ML models. Also, agnostic IML methods can potentially offer local interpretability that enables local predictions for a single instance.

Our future plan consists of applying agnostic IML methods for sector-based prediction problems, analyzing the results of agnostic IML methods for deep learning models, and improving the reliability of the results by evaluating consistency with assessments of domain experts.

6 Author contribution statements

In the scope of this study, ... conceptualized the idea, carried out the methodology and software development, performed experimental analysis; ... evaluated the results, writed visualized, reviewed; ... supervised and edited the article.

7 Ethics committee approval and conflict of interest statement

There is no need to obtain permission from the ethics committee for the article prepared. There is no conflict of interest with any person / institution in the article prepared.

8 References

[1] Morocho-Cayamcela ME, Lee H, Lim W. "Machine learning for 5g/b5g mobile and wireless communications: Potential, limitations, and future directions". *IEEE Access,* 7, 137184–137206, 2019.

[2] Baryannis G, Dani S, Antoniou G. "Predicting supply chain risks using machine learning: The trade-off between performance and interpretability". *Future Generation Computer Systems,* 101, 993–1004, 2019.

[3] Mori T, Uchihira N. "Balancing the trade-off between accuracy and interpretability in software defect prediction". *Empirical Software Engineering*, 24(2), 779–825, 2019.

[4] Doshi-Velez F, Kim B. "Towards a rigorous science of interpretable machine learning". *arXiv preprint arXiv:1702.08608,* 2017.

[5] Lundberg S, Lee SI. "A unified approach to interpreting model predictions". *arXiv preprint arXiv:1705.07874,* 2017.

[6] Fan A, Jernite Y, Perez E, Grangier D, Weston J, Auli M. "ELI5: Long form question answering*". arXiv preprint arXiv:1907.09190*, 2019.

[7] Ribeiro MT, Singh S, Guestrin C. "" Why Should İ Trust You?" explaining the predictions of any classifier". *in Proceedings of the 22nd ACM SIGKDD International Conference On Knowledge Discovery and Data Mining*, San Francisco, California, USA, 13-17 August 2016.

[8] Zhao L, Dong X. "An industrial internet of things feature selection method based on potential entropy evaluation criteria". *IEEE Access*, 6, 4608–4617, 2018.

[9] Jiang T, Gradus JL, Rosellini AJ. "Supervised machine learning: a brief primer". Behavior Therapy, 51(5), 675-687, 2020.

[10] Jovic´ A, Brkic´ K, Bogunovic´ N. "A review of feature selection methods with applications". in 2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO), Opajita, Croatia, 25-29 May 2015.

[11] Alelyani S, Tang J, Liu H. "Feature selection for clustering: A review". Editors: Aggarwal C, Reddy R. Data Clustering: Algorithms and Applications, 2013.

[12] Dy JG, Brodley CE. "Feature selection for unsupervised learning". *Journal of machine learning research*, 5, 845–889, 2004.

[13] Ang JC, Mirzal A, Haron H, and Hamed HNA. "Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection". *IEEE/ACM transactions on computational biology and bioin- formatics*, 13(5), pp. 971–989, 2015.

[14] Yıldırım S, Yıldız T. "A comparative analysis of text classification for Turkish language". *Pamukkale University Journal of Engineering Sciences*, 24(5), 879-886, 2018.

[15] Behura A. "The cluster analysis and feature selection: Perspective of machine learning and image processing". Editors: Satpathy R, Choudhury T, Satpathy S, Mohanty S N, Zhang X. Data Analytics in Bioinformatics: A Machine Learning Perspective, 249–280, John Wiley & Sons, 2021.

[16] Li X, Yi P, Wei W, Jiang Y, Tian L. "LNNLS-KH: A Feature Selection Method For Network İntrusion Detection". *Security and Communication Networks,* 2021, 1-22, 2021.

[17] Selvalakshmi B, Subramaniam M. "Intelligent ontology based semantic information retrieval using feature selection and classification". *Cluster Computing*, 22(5), 12871–12881, 2019.

[18] Du X, Li W, Ruan S, Li L. "Cus-heterogeneous ensemble-based financial distress prediction for imbalanced dataset with ensemble feature selection". *Applied Soft Computing*, 97, p. 106758, 2020.

[19] Yousef M, Kumar A, Bakir-Gungor B. "Application of biological domain knowledge based feature selection on gene expression data". *Entropy*, 23(1), 1-15,2021.

[20] Akalın F, Yumusak N. "Classification of acute leukaemias with a hybrid use of feature selection algorithms and deep learning-based architectures". *Pamukkale University Journal of Engineering Sciences*, 2022.

[21] Mohammadi S, Mirvaziri H, Ghazizadeh-Ahsaee M, Karimipour H. "Cyber intrusion detection by combined feature selection algo- rithm". *Journal of information security and applications,* 44, 80–88, 2019.

[22] Sharif M, Khan MA, Iqbal Z, Azam MF, Lali MIU, Javed MY. "Detection and classification of citrus diseases in agriculture based on optimized weighted segmentation and feature selection". *Computers and Electronics in Agriculture*, 150, 220–234, 2018.

[23] Rodriguez-Galiano VF, Luque-Espinar JA, Chica-Olmo M, Mendes MP. "Feature selection approaches for predictive modelling of groundwater nitrate pollution: An evaluation of filters, embedded and wrapper methods". *Science of the Total Environment*, 624, 661– 672, 2018.

[24] Xue B, Zhang M, Browne WN, Yao X. "A survey on evolutionary computation approaches to feature selection". *IEEE Transactions on Evolutionary Computation*, 20(4), 606–626, 2015.

[25] Sharma S, Kaur P. "A comprehensive analysis of nature-inspired meta-heuristic techniques for feature selection problem*". Archives of Computational Methods in Engineering*, 28(3), 1103-1127, 2021.

[26] Kumar V, Minz S. "Feature Selection: A Literature Review". *SmartCR*, 4(3), 211–229, 2014.

[27] Bolo´n-Canedo V, Sa´nchez-Maron˜o N, Alonso-Betanzos A, "A review of feature selection methods on synthetic data". *Knowledge and Information Systems*, 34(3), 483–519, 2013.

[28] Li H, Li CJ, Wu XJ, Sun J. "Statistics-based wrapper for feature selection: An implementation on financial distress identification with support vector machine". *Applied Soft Computing*, 19, 57– 67, 2014.

[29] Cui L, Bai L, Wang Y, Jin X, Hancock ER. "Internet financing credit risk evaluation using multiple structural interacting elastic net feature selection". *Pattern Recognition*, 114, 1-13, 2021.

[30] Jadhav S, He H, Jenkins K. "Information gain directed genetic algorithm wrapper feature selection for credit rating". *Applied Soft Computing*, 69, 541–553, 2018.

[31] Liang D, Tsai CF, Wu HT. "The effect of feature selection on financial distress prediction". Knowledge-Based Systems, 73, 289–297, 2015.

[32] Sivasankar E, Selvi C, Mahalakshmi S. "Rough set-based feature selection for credit risk prediction using weight-adjusted boosting en- semble method". *Soft Computing,* 24(6), 3975–3988, 2020.

[33] Zhang X, Hu Y, Xie K, Wang S, Ngai E, Liu M. "A causal feature selection algorithm for stock prediction modeling". *Neurocomputing,* 142, 48–59, 2014.

[34] Lin WC, Lu YH, Tsai CF. "Feature selection in single and ensemble learning-based bankruptcy prediction models". *Expert Systems*, 36(1), 1-8, 2019.

[35] Adadi A, Berrada M. "Peeking inside the black-box: a survey on explainable artificial intelligence (xai)". *IEEE Access*, 6, 52138– 52160, 2018.

[36] Yıldırım-Okay F, Yıldırım M, and Ozdemir S. "Interpretable machine learning: A case study of healthcare". *in 2021 International Symposium on Networks, Computers and Communications (ISNCC)*. IEEE, Dubai, United Arab Emirates, 31 October-2 November 2021.
[37] ElShawi R, Sherif Y, Al-Mallah M, Sakr S. "Interpretability in healthcare: A comparative study of local machine learning interpretability techniques". *Computational Intelligence*, 37(4), 1633-1650, 2021.
[38] Zablocki E, Ben-Younes H, Pe´rez P, Cord M. "Explainability of vision-based autonomous driving systems: Review and challenges". *arXiv preprint arXiv:2101.05307*, 2021.
[39] Demajo LM, Vella V, Dingli A. "Explainable AI for interpretable credit scoring". *arXiv preprint arXiv:2012.03749*, 2020.
[40] Bussmann N, Giudici P, Marinelli D, Papenbrock J. "Explainable AI in Fintech Risk Management". *Frontiers in Artificial Intelligence,* 3(26), 1-5, 2020.
[41] Ito T, Sakaji H, Izumi K, Tsubouchi K, Yamashita T. "GINN: Gradient Interpretable Neural Networks For Visualizing Financial Texts". *International Journal of Data Science and Analytics*, 9(4), 431–445, 2020.
[42] Cong LW, Tang K, Wang J, Zhang Y. "Alphaportfolio for investment and economically interpretable AI". *SSRN*, 2020.
[43] Grath RM, Costabello L, Van CL, Sweeney P, Kamiab F, Shen Z, Lecue F. "Interpretable credit application predictions with counterfactual explanations". *arXiv preprint arXiv:1811.05245*, 2018.
[44] Ghosh I, Dragan P. "Can financial stress be anticipated and explained? Uncovering the hidden pattern using EEMD-LSTM, EEMD-prophet, and XAI methodologies". Expert Systems with Applications, 181, 115026, 2022.
[45] Babaei G, Giudici P. "Which SME is worth an investment? An explainable machine learning approach. An explainable machine learning approach", 2021.
[46] Tra KL, Le HA, Nguyen TH, Nguyen, DT. "Explainable machine learning for financial distress prediction: evidence from Vietnam. Data", 7(11), 160, 2022.
[47] Ariza-Garzón MJ, Arroyo J, Caparrini A, & Segovia-Vargas MJ. "Explainability of a machine learning granting scoring model in peer-to-peer lending". IEEE Access, 8, 64873-64890, 2020.
[48] Misheva BH, Osterrieder J, Hirsa A, Kulkarni O, Lin, SF. "Explainable AI in credit risk management", arXiv preprint arXiv:2103.00949, 2021.
[49] Bracke P, Datta A, Jung C, Sen, S. "Machine learning explainability in finance: an application to default risk analysis", SSRN: Amsterdam, The Netherlands, 2019.
[50] Chandrashekar G, Sahin F. "A survey on feature selection methods". *Computers & Electrical Engineering*, 40(1), 16–28, 2014.
[51] Li J, Cheng K, Wang S, Morstatter F, Trevino RP, Tang J, Liu H. "Feature selection: A data perspective". *ACM Computing Surveys (CSUR*), 50(6),1–45, 2017.
[52] Saeys Y, Inza I, Larranaga P. "A review of feature selection techniques in bioinformatics". *Bioinformatics*, 23(19), 2507– 2517, 2007.
[53] Carvalho DV, Pereira EM, Cardoso JS. "Machine learning interpretability: A survey on methods and metrics". *Electronics*, 8(8), 1-34, 2019.
[54] Vellido A, Mart´ın-Guerrero JD, Lisboa PJ. "Making machine learning models interpretable." in European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), Bruge, Belgium, 25-27 April 2012.
[55] Molnar C, Interpretable machine learning. Lulu. com, 2020.
[56] Ruping S. "Learning interpretable models". 2006.
[57] Robnik-S˘ikonja M, Bohanec M. "Perturbation-based explanations of prediction models". *in Human and machine learning*. Springer, 2018.
[58] Mothilal RK, Sharma A, Tan C. "Explaining machine learning classifiers through diverse counterfactual explanations". *in Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, NY, USA, 37-30 January 2020.
[59] Zikeba M, Tomczak SK, Tomczak JM. "Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction". *Expert Systems with Applications*, 58, 93-101, 2016.
[60] Real R, Vargas JM. "The probabilistic basis of jaccard's index of similarity". *Systematic Biology*, 45(3), 380–385, 1996.
[61] Tata S, Patel JM. "Estimating the selectivity of tf-idf based cosine similarity predicates", ACM Sigmod Record, 36(2), 7-12, 2007.