IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 5 December 2024, accepted 12 December 2024, date of publication 23 December 2024,
date of current version 30 December 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3521242

RESEARCH ARTICLE

Balancing Predictive Performance and Interpretability in Machine Learning: A Scoring System and an Empirical Study in Traffic Prediction

FABIAN OBSTER[1,2], MONICA I. CIOLACU[3], AND ANDREAS HUMPE[4], (Member, IEEE)
[1]Department of Business Administration, University of the Bundeswehr Munich, 85577 Neubiberg, Germany
[2]Department of Statistics, LMU Munich, 80539 Munich, Germany
[3]Faculty of Social and Educational Sciences, University of Passau, 94032 Passau, Germany
[4]Institute for Applications of Machine Learning and Intelligent Systems (IAMLIS), Munich University of Applied Sciences, 80335 Munich, Germany

Corresponding author: Fabian Obster (fabian.obster@unibw.de)

This work was supported by the Digitalization and Technology Research Center of the Bundeswehr (dtec.bw) funded by European Union–NextGenerationEU.

ABSTRACT This paper investigates the empirical relationship between predictive performance, often called predictive power, and interpretability of various Machine Learning algorithms, focusing on bicycle traffic data from four cities. As Machine Learning algorithms become increasingly embedded in decision-making processes, particularly for traffic management and other high-level commitment applications, concerns regarding the transparency and trustworthiness of complex 'black-box' models have grown. Theoretical assertions often propose a trade-off between model complexity (predictive performance) and transparency (interpretability); however, empirical evidence supporting this claim is limited and inconsistent. To address this gap, we introduce a novel interpretability scoring system - a Machine Learning Interpretability Rank-based scale - that combines objective measures such as the number of model parameters with subjective interpretability rankings across different model types. This comprehensive methodology includes stratified sampling, model tuning, and a two-step ranking system to operationalize this trade-off. Results reveal a significant negative correlation between interpretability and predictive performance for intrinsically interpretable models, reinforcing the notion of a trade-off. However, this relationship does not hold for black-box models, suggesting that for these algorithms, predictive performance can be prioritized over interpretability. This study contributes to the ongoing discourse on explainable AI, providing practical insights and tools to help researchers and practitioners achieve a balance between model complexity and transparency. We recommend to prioritise more interpretable models when predictive performance is comparable. Our scale provides a transparent and efficient framework for implementing this heuristic and improving parameter optimization. Further research should extend this analysis to unstructured data, explore different interpretability methods, and develop new metrics for evaluating the trade-off across diverse contexts.

INDEX TERMS Explainable AI, predictive analysis, interpretability scoring, empirical analysis, bicycle traffic data, bias-variance, trade-off.

I. INTRODUCTION

In recent years, the use of Machine Learning for decision-making and forecasting has grown rapidly and gained in popularity. Artificial Intelligence (AI) solutions,

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

which often leverage Machine Learning techniques, are expected to experience significant growth. According to the International Data Corporation (IDC), the global AI market will grow from 235 billion dollars to 631 billion by 2028 [1]. Despite the economic challenges caused by the COVID-19 pandemic, the Russian war against Ukraine, and the green transition, companies see significant value in AI-enhanced products and services. The emergence of AI is a major inflection point in the technology industry. Companies will race to introduce AI-enhanced products and services, allocate over 40 percent of their core IT spend to AI initiatives, and invest in employee training [2]. Gartner identified the three significant strategic technology trends for 2024. These are the importance of protecting investment, the rise of the builders, and the need to deliver the value [3]. The IBM Global AI adoption index findings indicate that 42 percent of enterprise-scale companies have already deployed AI in their business. Furthermore, 59 percent of those already exploring deploying AI have accelerated their roll-out or investments in technology. IBM identifies top barriers in preventing AI deployment as a lack of AI skills and expertise (33 percent), the complexity of data 25 percent and ethical concerns (23 percent). Business adoption is crucial, trustworthy, and explainable AI is critical, and the key to increasing AI adoption is the ability to access data anywhere [4].

Much of the focus has been on improving predictive performance, with less emphasis on model interpretability. While models like deep learning and random forests offer high predictive performance, they often function as 'black-boxes,' making their internal decision processes difficult to understand [5]. On the other hand, simpler models like decision trees and linear models offer more interpretability but may fall short in terms of performance [6]. The current gap in research lies in balancing these two aspects—predictive performance and interpretability—and ensuring that highly interpretable models are not dismissed in favor of more complex, opaque models. Our study seeks to fill this gap by operationalizing interpretability and applying it to real-world data, demonstrating how the trade-off between these aspects can be evaluated and balanced. This trade-off has been a recurring theme in the literature, with many researchers suggesting that more complex models generally outperform simpler ones in terms of predictive performance but are harder to interpret [7], [8]. Theoretically, it has been argued that more complex models are more accurate [6], [9]. Often, a linear relationship is hypothesized between performance/accuracy and explainability. Table 1 shows some studies that suggest such a relationship.

Although the theoretical trade-off between performance and explainability/interpretability is well documented, empirical evidence is largely lacking and contradictory. Interpretability between model types based on subjective rankings has been linked to predictive performance, such as in [7] focusing on fairness-aware models and mainly decision trees, or in [14], also considering unstructured data in contrast to others. Another strain of research focuses on a specific class of Machine Learning models only, such as deep learning [15]. However, a unified framework for evaluating interpretability for a variety of model types beyond subjective ratings is lacking in the literature.

TABLE 1. Overview of literature.

Authors	In order of performance (low to high) and explainability (high to low)
Jo et al. (2023) [7]	Decision trees, Linear regression, K-nearest neighbors, SVMs, Random Forest, Deep Learning
Arrieta et al. (2020) [10]	Rule-based learning, Linear / Logistic Regression, Decision Tree, kNN, Generalized Additive Models, Bayesian Models, SVM, Ensembles, Deep Learning
Yang et al. (2019) [11]	Linear regression, Decision trees, K-nearest neighbors, Random forests, SVMs, Deep neural networks
Gunning et al. (2019) [12]	Decision trees, Bayesian Belief Nets, SVMs, Random Forests, Deep Learning
Dam et al. (2018) [13]	Decision Trees and Classification Rules, Graphical Models (e.g. Bayesian Networks), SVMs, Ensemble Methods (e.g., Random Forests), Deep Learning (Neural Networks)

For high level of commitment decisions, such as those in criminal justice, military defence or medicine, the consequences of using black-box models that produce results that are not explainable are well documented in the literature [6]. Intelligent systems based on machine learning algorithms often require trust, must avoid biases, and comply with regulations and policies. Hence, good forecasting performance alone is not sufficient for selecting the preferred algorithm. However, we argue that explainability and interpretability are also crucial for non-high-stake decisions and sometimes even more so to some degree than forecasting performance. For instance, in policy decisions, human lives or freedoms may not be at risk. However, these decisions could involve substantial financial funds or have long-lasting, sometimes irreversible effects (e.g., building a highway or large bridge). Whereas, in a similar setting of short-term traffic management, forecasting performance might also be more important than explainability and interpretability. Besides high and low-stake decisions, strategic (long-term) and tactical (short-term) decisions play a key role in weighing performance against explainability and interpretability.

For the empirical analysis, we choose a traffic management example with strategic and short-term dimensions to empirically verify the hypothesized relationship between performance and explainability/interpretability of various machine learning algorithms. The data set consists of structured data and the conclusion cannot be generalized to unstructured or other data sets. The paper aims to operationalize interpretability within a unified framework including subjective and objective components, verify its relationship with predictive performance, and potentially propose an alternative view of the hypothesized trade-off between predictive performance and interpretability/explainability found in the literature. Through the interpretability score, one can compare the

interpretability of two Machine Learning models with a specific set of parameters, regardless of the algorithm used for the fitting or type of Machine Learning model. We believe that without such a framework, the interpretability-predictive performance tradeoff is a fuzzy concept, and the ongoing debate lacks a conceptual body enabling the transferability of specific results and consistently finding a balance between performance and interpretability. The paper is structured as follows. Section II reviews and discusses machine learning, artificial intelligence, predictive performance, interpretability and explainability. Section III describes the data, data collection and preparation process, detailing the data basis for the analysis. Section IV presents the methodology, outlining the statistical tools and procedures employed to examine the data and test the proposed hypotheses. Section V reports the analysis results, while Section VI deals with the study limitations and gives recommendations for future research. Finally, section VII discusses the main findings and Section VIII shows the overall conclusions based on the results.

II. ARTIFICIAL INTELLIGENCE AND EXPLAINABLE AI

The past decade has seen rapid advancements and increasing use of artificial intelligence (AI) in industry. This leap in performance has often been achieved through high model complexity with ''black-box'' approaches, leading to uncertainty about how these models operate and arrive at decisions. This is highly problematic, especially in sensitive and critical areas such as autonomous driving or healthcare. Consequently, scientific interest in explainable artificial intelligence (XAI) has grown. This field focuses on developing new methods to explain and interpret machine learning models, contrasting with established methods that often prioritize accuracy over interpretability.

Explainability and interpretability, while related and sometimes used synonymously, are distinct concepts in the context of AI. Interpretability refers to the extent to which a cause and effect can be observed within a model, whereas explainability involves making the internal mechanisms of a model understandable to humans. For the purposes of this paper, we will use these definitions to ensure clarity.

For an overview of the development of explainable artificial intelligence (XAI) see Linardatos et al. [16] and Confalonieri et al. [17]. Various XAI methods have been developed, including local explanation methods that approximate individual predictions of a black-box model using local surrogate models. One such method is the LIME (Local Interpretable Model-agnostic Explanations) algorithm [18]. Here, the LIME approach exploits the fact that the trained black-box model can be queried multiple times for the predictions of specific instances. By changing the data used for training, LIME generates a new data set. After the black-box model is fed the modified data, it creates a new interpretable model from the predictions generated over the new data set. When an XAI method provides an explanation for only a particular instance, it is called a local approach, and when the method explains the entire model, it is called global. While LIME is a local approach, global XAI approaches such as PDPbox or Shapley Additive Explanations (SHAP) [19] are also regularly used in different applications [20]. Doshi-Velez and Kim [21] propose a taxonomy for interpretable machine learning methods and emphasize the need for a rigorous science of interpretability. They categorize interpretability methods into three types: transparency, post-hoc explanations, and intrinsically interpretable models. Gilpin et al. [22] provide a comprehensive overview of interpretability in machine learning, discussing various methods and their applications. They highlight the importance of model transparency and the challenges associated with interpreting complex models. Tjoa and Guan [23] survey the current state of explainable AI, particularly in the medical field. They discuss the necessity for explainability in medical AI systems to ensure trust and compliance with regulatory standards. Carvalho et al. [24] survey methods and metrics for machine learning interpretability, categorizing them into intrinsic and post-hoc techniques. They emphasize the importance of selecting appropriate methods based on the specific application and requirements. However, the use of interpretable algorithms to explain non-interpretable algorithms raises questions. If interpretable results are necessary to explain black boxes, why not use white-box models from the beginning? Our results provide both, theoretical insights into the generally hypothesized trade-off between predictive performance and explainability of interpretable and non-interpretable machine learning algorithms in a real-world example, e.g., traffic prediction for urban planning, as well as a practical benefit through the extension of purely predictive models to explanatory models which can also be used prescriptively. This study thus seeks to explore the relationship between the interpretability of machine learning models and their predictive capabilities. To achieve this goal, the following research questions are addressed:

- RQ 1: Is there a clear relationship between the interpretability ranking of machine learning models or algorithms and their predictive performance?
- RQ 2: Is the model type rank associated with predictive performance?
- RQ 3: Is the number of model parameters associated with predictive performance?
- RQ 4: Is interpretability associated with predictive performance within model types?

Answering these research questions will help clarify the relationship between model interpretability and predictive performance, guiding the development and selection of algorithms for applications where both are crucial. The examination of model type rank and its association with predictive performance will further enhance our understanding of how different algorithms perform in practice, providing a comprehensive framework for evaluating machine learning models.

TABLE 2. Data sources.

Sources	Bike counting data	Weather data	Public holidays
Berlin	https://www.berlin.de/sen/uvk/verkehr/verkehrsplanung/radverkehr/weitere-radinfrastruktur/zaehlstellen-und-fahrradbarometer/#dauer	https://opendata.dwd.de	https://kalender-2017.net, https://kalender-2018.net, https://kalender-2019.net,https://kalender-2020.net
Dusseldorf	https://opendata.duesseldorf.de/dataset/jahres%C3%BCbersicht-der-dauerz%C3%A4hlstellen-radverkehr-seit-2012	https://opendata.dwd.de	https://kalender-2017.net, https://kalender-2018.net, https://kalender-2019.net,https://kalender-2020.net
Munich	https://www.opengov-muenchen.de/pages/raddauerzaehlstellen	https://opendata.dwd.de	https://kalender-2017.net, https://kalender-2018.net, https://kalender-2019.net,https://kalender-2020.net
Vienna	https://www.data.gv.at/katalog/dataset/stadt-wien_radverkehrs-zhlungenderstadt-wien/resource/c2d89b4e-8193-4615-b477-67a68c488af3#resources	https://opendata.duesseldorf.de/dataset/jahres%C3%BCbersicht-der-dauerz%C3%A4hlstellen-radverkehr-seit-2012	https://www.ferienwiki.at/feiertage/2017/at, https://www.ferienwiki.at/feiertage/2018/at, https://www.ferienwiki.at/feiertage/2019/at, https://www.ferienwiki.at/feiertage/2020/at

III. RESEARCH DESIGN

The analysis used data from bicycle counters, which are electronic devices that record the number of bicycles and are used to register bicycle traffic permanently and automatically. They are sometimes called bicycle barometers and are used in many cities today [25]. To utilize a multi-year data set, four cities with permanent automatic bicycle counting stations operational since at least 2017 were selected. In addition to daily bicycle counts as the dependent variable, weather data and dummy variables for public holidays, lockdowns, and the introduction of pop-up bike lanes were collected as independent variables. Table 1 gives an overview of the sources for bike traffic, weather data, and public holidays. For lockdowns and the introduction of pop-up lanes, we had to rely on local media and newspaper sources.

IV. METHODOLOGY

We conducted a comprehensive data-driven analysis to empirically verify the hypothesized relationship between the performance and interpretability of various Machine Learning algorithms based on rankings shown in Table 1. Finally, we selected a bicycle traffic dataset that includes both strategic and short-term dimensions. The dataset comprises structured data, and therefore, our conclusions should not be generalized to unstructured data or other datasets. Our objective is to verify the relationship and potentially offer an alternative view of predictive performance versus interpretability trade-off discussed in the literature.

The data and code for the analysis and the generation of the figures and table can be found on GitHub (https://github.com/FabianObster/xai_bikes). An overview of the used data can be found in Table 2.

The design of our analysis involves using 50 Percent of the data for model training, with the remaining 50 Percent reserved for model evaluation. The data were randomly assigned to either the training or test set for each year. This stratified sampling method was utilized to ensure that the training and test sets were comparable concerning their temporal distribution. Models were fitted for each of the four cities. For each model type, we fitted three distinct models: one with a relatively low number of parameters, another with a relatively high number of parameters, and a third model that was tuned using 10-fold cross-validation on the training dataset. We estimated the resulting model parameters using the full dataset of training data to facilitate both between- and within-model-type comparisons of machine learning models.

Different machine learning algorithms require a different amount of hyperparameters. To ensure comparability, 25 distinct hyperparameter combinations were consistently tuned for each model type's tuned version. If no hyperparameters existed, no tuning was performed, which happened for the linear model. The loss function for all machine learning algorithms was set to the L^2 loss. For measuring predictive performance, we used the root mean squared error (RMSE)

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{n}},$$

the mean absolute value (MAE)

$$\text{MAE} = \frac{\sum_{i=1}^{n}|y_i - \widehat{y}_i|}{n},$$

and the pearson correlation between the predicted values and the actual values (R^2)

$$R^2 = \frac{\text{cov}[y, \widehat{y}]}{\sigma_y \sigma_{\widehat{y}}} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\widehat{y}_i - \bar{\widehat{y}})}{\sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}{n}}\sqrt{\frac{\sum_{i=1}^{n}(\widehat{y}_i - \bar{\widehat{y}})^2}{n}}}$$

in the test dataset. The RMSE evaluates the overall predictive performance of the models by measuring the average squared magnitude of the errors. As the magnitudes are squared, predictive outliers have a strong effect on the overall metric, we also present the MAE which averages the magnitude of errors. Lower values for both metrics indicate higher predictive performance. Both these metrics are unbounded and are therefore hard to compare if outcomes of different scales are considered. The R^2 on the other hand can only take values between zero and one increasing the comparability and interpretability of the results. Here, higher values indicate higher predictive performance.

TABLE 3. Overview over model architecture, used hyperparameters and interpretability scale.

Model type	Model type Abbreviation	Tuned hyperparameters	Software	Version 1	Version 2	Interpretability Rank
Linear Model	lm	-	Base R	-	-	1
Regression tree	Tree	cost_complexity, tree_depth	Tree [26], part [27]	mincut=3, minsize=6, mindev=0.005	mincut=1, minsize=2, mindev=0.0005	2-4
k-nearest neighbors	knn	neighbors	FNN [28], kknn [29]	K=27	K=1	5-7
Boosted generalized additive models	mb	mstop	mboost [30]	Mstop=2000	Mstop=20000	8-10
Boosted glms with interactions	mbsp	mstop	mboost [30]	Mstop=1500	Mstop=15000	11-13
Support vector machines	svm	Cost, rbf_sigma	e1071 [31], kernlab [32]	Kernel=polynomial, degree=3	Kernel=radial, degree=3	14-16
Boosted trees	gbm	Ntree, interaction.depth	gbm [33]	Ntree=100, interaction.depth=3, n.minobsinnode=1	Ntree=100, Interaction.depth=7, n.minobsinnode=1	17-19
Random Forest	rf	ntree, mtry	randomForest [34]	ntree=500, mtry=3	ntree=1000 mtry=6	20-22
Neural network	nn	Hidden_units, penalty (activation=relu)	neuralnet [35], nnet [36]	First hidden layer=5, activation = logistic	First hidden layer=5, second hidden layer=3, activation = logistic	23-25

All evaluation metrics were computed per city and per year. Table 3 provides an overview of the model types, tuned hyperparameters, software packages used, and an interpretability rank. To measure interpretability, we followed a two-step approach. In the first step, each model type (e.g., support vector machines, linear models, neural networks) was assigned an interpretability rank based on a predefined scale [10]. In the second step, the complexity of interpretation was assessed within each model type. For instance, a linear model with a single covariate is more straightforward to interpret than one with a thousand covariates. Thus, the complexity of each model was characterized by the number of parameters it contained. Additionally, a greater number of parameters increases the time required by a human to understand the underlying functional dependencies.

The first step captured the complexity of the functional representation, while the second step allowed us to differentiate between the number of such transformations defined by the first step. To synthesize a ranked scale, models were first sorted according to their type-based rank and subsequently, within each rank, by their complexity-based rank. This process resulted in an ordinal ranking system, facilitating comparison across a broad spectrum of machine learning models. However, this ranking should be applied to a predetermined set of models. Should an additional model be introduced, the ranks of all models previously positioned higher will increase by one.

The interpretability rank only describes the final model output, not the way a machine learning algorithm learns. In this sense, there is no differentiation between for example bagged trees (random forest) and boosted trees (adaboost/xgboost) if the resulting model consists of a linear combination of trees, even though in the machine learning context both are considered different algorithms. In our evaluation of the association metrics between the interpretability scale $(s_i)_{i\leq n}$ and the predictive measure $(p_i)_{i\leq n}$ converted to ranks $R(s)$ and $R(p)$, we employed the Spearman rank correlation coefficient ρ_s,

$$\rho_s = \frac{\mathrm{cov}[R(s), R(p)]}{\sigma_{R(s)}\sigma_{R(p)}},$$

using the same notation as in the Pearson correlation. The Spearman rank Coefficient also takes values between zero and one, with higher values indicating a stronger association. It is a measure of the monotonicity of the relation between two variables, meaning that monotonous transformations of the underlying data do not change the value, making it less prone to outliers. To detect univariate associations between the interpretability scale and predictive performance, Pearson correlation coefficients were used to examine the relationship between the number of parameters and predictive performance. To identify the within-model-type interpretability effect, we applied linear mixed-effects regression with model type as a random effect, simple linear regression was used to determine the impact of the number of parameters and the interpretability scale. We used the R package 'mgcv' [37] for fitting the regression models, 'ggplot2' [38] for visualizations, and 'dplyr' [39] for data manipulation. Having established our methodology, we now present the empirical results that shed light on the relationship between model interpretability and predictive performance. These results not only validate our hypotheses but also offer insights into the nuances of different machine learning models.

V. RESULTS

Following the methodologies outlined in the previous section, this results section now reports the observed associations. For the results of the correlation analysis between parameters and interpretability see the table below:

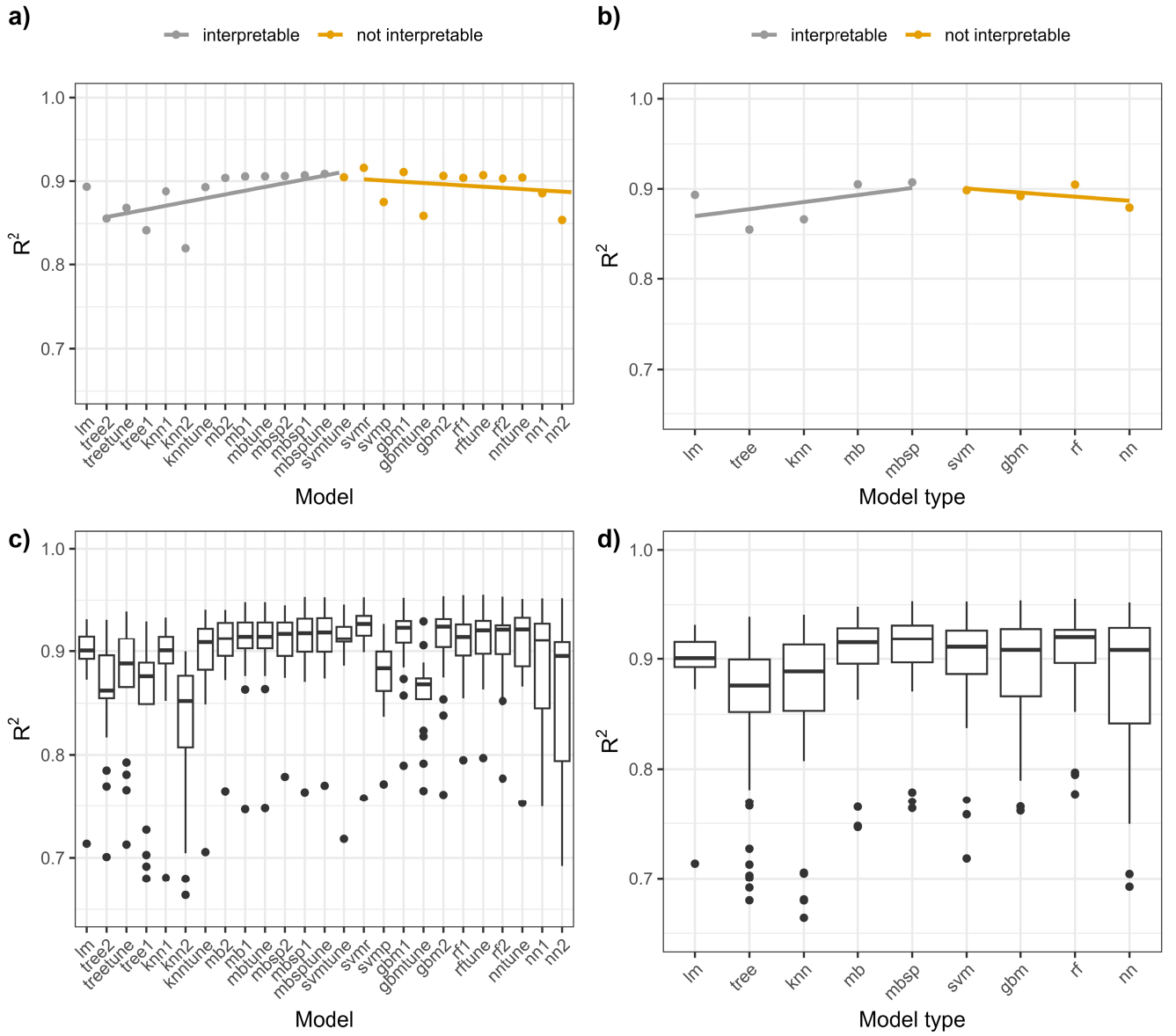

FIGURE 1. a) and b) Scatterplot Average R^2 of each model left (a)) and boxplot of R^2 for each model type (b)). Stratified linear trend by intrinsically interpretability indicated by color. c) and d) Boxplot showing R^2 for each model (c)) and each model type (d)). The x-axes of all figures is arranged by the interpretability scale. When model tuning was performed, the rank depended on the actual number of estimated parameters.

Table 4 shows the correlation between the number of parameters and predictive performance, indicating that for interpretable machine learning algorithms, predictive performance increases with more complex models based on the interpretability scale. However, predictive performance does generally not increase with the interpretability scale for generally non-interpretable machine learning algorithms.

Based on the correlation results between the number of parameters and predictive performance of the different models, the number of parameters seems to be positively associated with predictive performance in all considered scenarios. Figure 1 below gives a graphical representation of the analysis.

With the results shown in Table 4 and Figure 1, we now analyze and interpret these findings within the framework of the initial research questions, examining the extent to which they might support our hypotheses.

A. RQ 1: THE INTERPRETABILITY SCALE

Overall, there is an association between the interpretability rank and predictive performance. However, the magnitude of the association depends on model tuning as shown in Table 4 and Figure 2. Considering the tuned as well as the untuned models, for the intrinsically interpretable models the rank correlation between the interpretability rank and

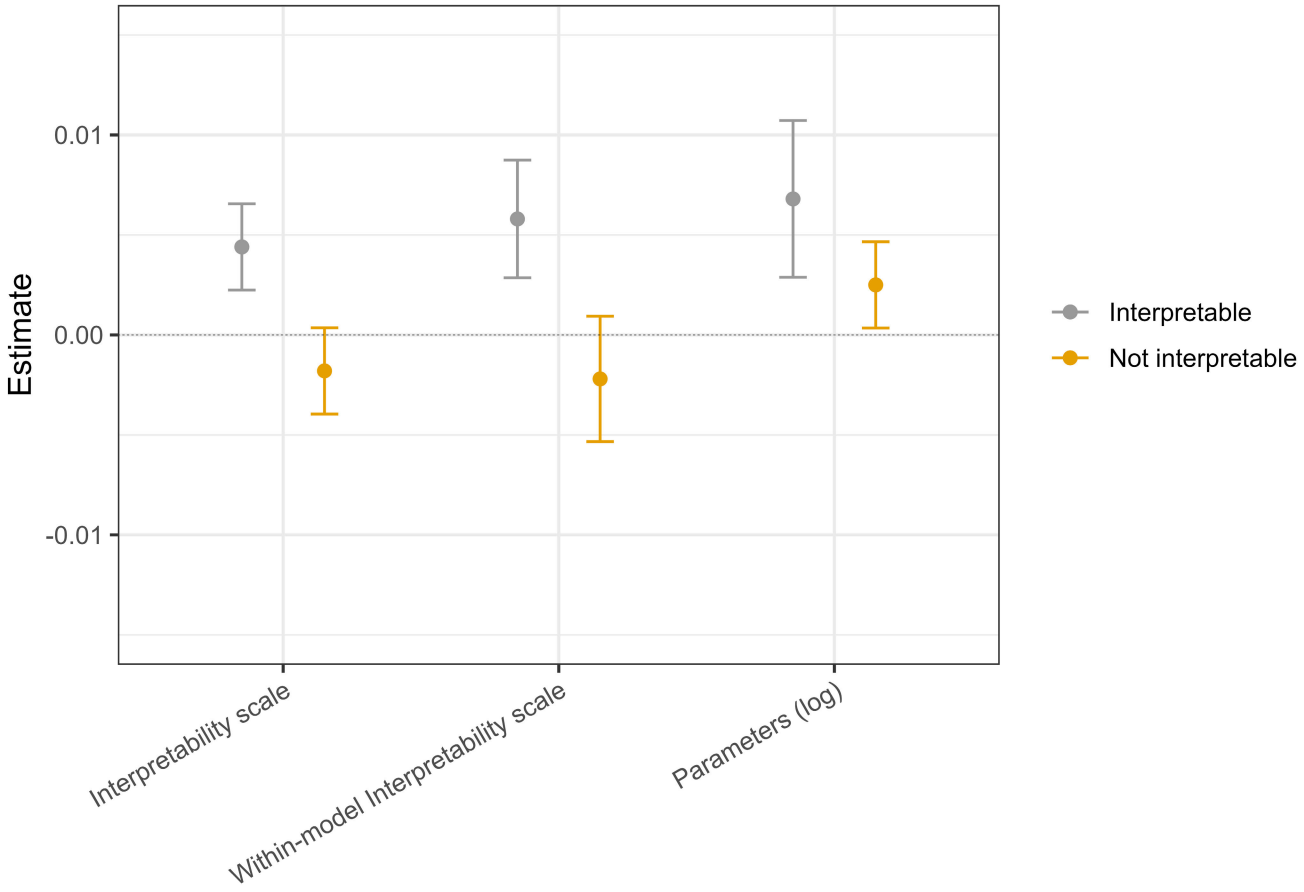

FIGURE 2. Estimates of the models quantifying the association between the Interpretability scale (left) and the logarithmic number of parameters (right) with the R^2 based on linear regression models. The within-model association between the interpretability scale and the R^2 based on the generalized mixed model with model type as a random effect is shown in the middle. Find the exact estimates and further metrics in Table 5.

TABLE 4. Association between parameters/interpretability with predictiveness (R^2). Pearson Correlation coefficient of the R^2 and the number of parameters (cor_params) as well as the logarithmic number of parameters (log_cor_params). Kendall tau (cor_rank) between R^2 and the interpretability scale.

subset tuning	Subset interpretable	Correlation params	Correlation log params	Correlation rank
All models (n=25)	Not Interpretable	0,128	0,161	-0,049
All models (n=25)	Interpretable	0,153	0,224	0,356
Tuned models (n=8)	Not interpretable	0,241	0,164	0,203
Tuned models (n=8)	Interpretable	0,125	0,271	0,379
All models (n=25)	All models	0.104	0.197	0.159
Tuned models (n=8)	All models	0.149	0.177	0.11

the R^2 was 0.36, and for the black box models −0.05. When considering only tuned models, the association was stronger, with rank correlations of 0.38 for interpretable models and 0.2 for black box models. Figure 1a illustrates the average R^2 for each model sorted by the interpretability scale showing a visible positive trend at the lower end of the scale including the intrinsically interpretable models. Futhermore, the observed trend is confirmed by the boxplot presented in Figure 1c. Figure 3 depicts the relationship between the interpretability scale and predictive performance including a local smoothed line highlighting the nonlinearity if the differentiation between black-box and intrinsically interpretable models is not considered. The findings of the descriptive analysis are confirmed by the results of the linear regression model depicted in Figure 2, which shows the error bars of the model fitted to the data using only the interpretable models and the non-interpretable models. There is a significant positive association for interpretable models which is not the case for non-interpretable models. The confidence intervals in Figure 2 are non-overlapping for the Interpretability scale and the within-model interpretability indicating a significant interaction (See documented code in GitHub https://github.com/FabianObster/xai_bikes). On average, the effect of the interpretability scale on the R^2 is higher by 0.006 ($p < 0.001$) per interpretability rank for

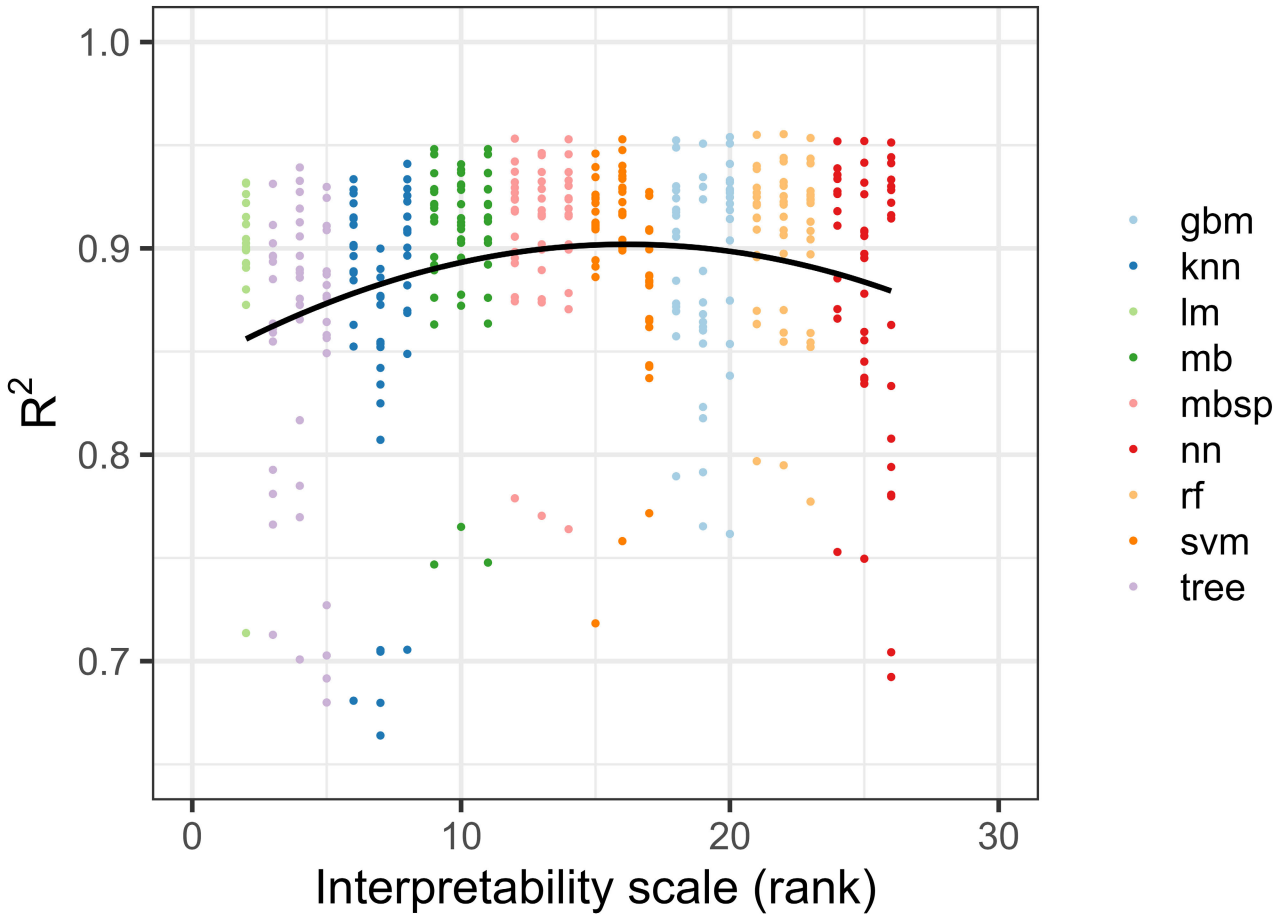

FIGURE 3. Scatterplot showing the interpretability scale on the x-axis vs the R^2 on the y-axis for all models. The color indicates the model type. The smoothed line is based on the LOESS estimator computed with ggplot2.

interpretable models compared to black-box models. For the within-model interpretability, the interaction effect is equal to 0.008 ($p < 0.001$). The logarithmic number of parameters does not show significant deviations between interpretable and black-box models ($p = 0.051$).

Next, we will analyze the two components contributing to the interpretability rank: the number of parameters and the qualitative rank of the model type as hypothesized in research questions two and three.

B. RQ 2: MODEL TYPE RANK

Referring to Figures 1b and 1d the trend, already shown on the model interpretability scale, is also visible in the model type interpretability rank. There is a positive trend with increasing model rank interpretability for intrinsically interpretable models and a slight negative trend for the black box models. Further metrics of predictive performance, the root mean squared error, and the mean absolute error, as shown in Table 6 also confirm the relationship already observed for the R^2. Tables 10 and 11 in the Appendix show the standard deviation of the median of the metrics of predictive performance across the model types.

C. RQ3: NUMBER OF PARAMETERS

We find a positive relationship between the number of parameters and predictive performance, as illustrated in Figure 4. This holds for both the entire set of models and the tuned models.

Given that the number of parameters differs largely across model types, with linear models having a maximum number of parameters limited by the number of variables, whereas models like random forests may have a possibly unbounded number of parameters (e.g., the number of trees), the analysis is conducted using the logarithm of the parameter count. Figure 4 shows a positive association between the logarithmic number of parameters and the R^2. This is the case when considering all models as well as only the tuned models. The Pearson correlation using all models is 0.2 and 0.18 for tuned models. The rank correlation is 0.16 and 0.11, respectively, showing a stronger difference. Note that for the interpretability scale, we are only interested in the rank association, which is invariant under monotonous transformations of the exact number of parameters. Table 7 summarizes the root mean squared error (RMSE), mean absolute error (MAE), number of parameters, and ranks

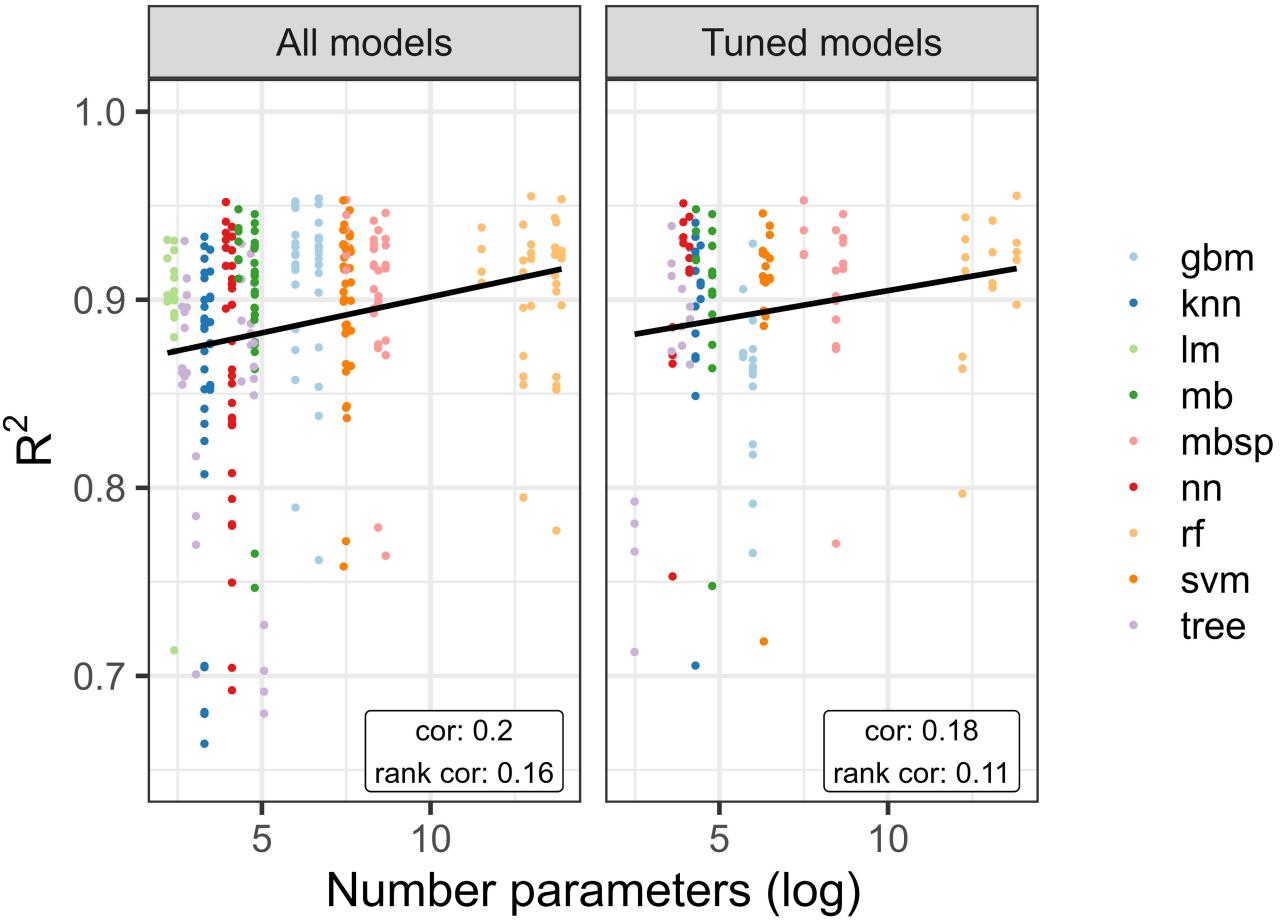

FIGURE 4. Scatterplot showing the logarithmic number of parameters on the x-axis vs the R^2 on the y-axis. The Left depicts all models including the tuned and untuned models and the right only the tuned model. The color indicates the model type.

averaged across the datasets for each model. Table 8 and Table 9 show the corresponding standard deviation and median for the same metrics. Figure 2 confirms the significant positive association between the logarithmic number and the predictive performance for both interpretable and non-interpretable models. The effect of the logarithmic number of parameters does not show significant deviations between interpretable and black-box models ($\widehat{\beta} = 0.004$, $p = 0.051$), yet interpretable models are associated with a higher effect (see documented code).

D. RQ 4: WITHIN-MODEL INTERPRETABILITY

Based on the linear regression with random effect model type shown in Figure 2 and Table 5, the effect in RQ 2 of the interpretability scale is confirmed. It can be concluded that the interpretability scale is associated with predictive performance for interpretable models beyond the type of machine learning model, which is not the case for non-interpretable models. The confidence intervals in Figure 2 are non-overlapping indicating a significant interaction, which is confirmed by the interaction analysis. On average, the effect of the within-model interpretability on the R^2 is higher by 0.008 ($p < 0.001$) per interpretability rank for interpretable models compared to black-box models.

In summary, the empirical analysis provides new insights into understanding the relationship between predictive performance and interpretability of machine learning algorithms. These findings establish a basis for a critical discussion on the limitations of the current research and the potential for future research.

VI. LIMITATIONS AND FUTURE WORK

Since there is no single best global model or model type that performs best for all datasets, the presented results should be viewed within the context of bike traffic in four cities.

The interpretability of machine learning model types is subjective and depends on the background of the person interpreting. Consequently, the hierarchy of model types presented in this article may not align with every individual's perception of interpretability. However, the scale can be adjusted in step one of the scoring procedures to reflect alternative rankings. This subjectivity might affect the scoring consistency across different evaluators. We addressed this issue by incorporating rankings proposed in the literature,

TABLE 5. Model results. Each linear (mixed) regression was fitted once for interpretable models and once for black-box models interpretable (Not interpretable) with the outcome R^2. Independent variables are the interpretability scale (rank) and the logarithmic number of parameters. For within-model interpretability the random effect indicating the model type (s(model type)) was used with the restricted maximum likelihood estimation.

Independent variable	Estimate	Std. Error	t value	p-value
Within-model Interpretability scale, Interpretable				
(Intercept)	0.834	0.014	59.518	<0.001
rank	0.006	0.002	3.794	<0.001
s(model type)				0.19
Within-model Interpretability scale, Not interpretable				
(Intercept)	0.940	0.033	28.531	<0.001
rank	-0.002	0.002	-1.414	0.159
s(model type)				0.090
Interpretability scale, Interpretable				
(Intercept)	0.849	0.009	90.167	<0.001
rank	0.004	0.001	4.150	<0.001
Interpretability scale, Not interpretable				
(Intercept)	0.930	0.022	42.515	<0.001
rank	-0.002	0.001	-1.674	<0.09
Parameters (log), Interpretable				
(Intercept)	0.851	0.011	80.987	<0.001
log(params)	0.007	0.002	3.395	<0.001
Parameters (log), Not interpretable				
(Intercept)	0.875	0.009	97.401	<0.001
log(params)	0.002	0.001	2.302	0.022

TABLE 6. Root mean squared error (RMSE), mean absolute error (mae), number of parameters, and rank averaged across the datasets for each model type.

Model type	Rmse	Mae	R^2	Parameters	Rank
gbm	8307.825	6909.936	0.892	525.49	19
knn	6501.046	5083.736	0.867	44.49	7
lm	6006.836	4761.446	0.893	10.529	2
mb	5627.202	4435.51	0.905	109.098	10
mbsp	5595.134	4423.797	0.907	4217.078	13
nn	7466.918	5870.733	0.879	56.404	25.1
rf	5664.8	4466.608	0.905	591816.3	22
svm	5879.4	4666.432	0.899	1414.765	16
tree	6606.812	5128.612	0.855	57.745	4

TABLE 7. root mean squared error (RMSE), mean absolute error (mae), number of parameters, and rank averaged across the datasets for each model.

Model	Rmse	Mae	R^2	Parameters	Rank
gbm1	5400.867	4252.405	0.911	400	18.2
gbm2	5465.016	4288.057	0.906	800	20
gbmtune	14057.59	12189.35	0.859	376.5	18.8
knn1	6011.299	4805.228	0.888	28.5	6
knn2	7693.028	5852.431	0.819	28.5	7
knntune	5798.812	4593.55	0.893	76.5	8
lm	6006.36	4761.446	0.893	10.5	2
mb1	5580.98	4380.744	0.906	109.2	9.2
mb2	5721.03	4545.42	0.904	108.9	9.8
mbsp1	5560.05	4383.754	0.907	4614.9	13
mbsp2	5678.33	4510.658	0.906	3690.6	12.2
mbsptune	5546.64	4376.98	0.909	4345.7	13.8
mbtune	5579.06	4380.365	0.906	109.18	11
nn1	7055.64	5679.49	0.886	58.6	24.5
nn2	10400.72	8032.075	0.854	58.6	25.5
nntune	4167.935	3294.45	0.904	50.5	25.4
rf1	5753.246	4566.58	0.904	313479.3	21.2
rf2	5635.654	4416.219	0.903	955009.5	23
rftune	5605.501	4417.025	0.907	506960.1	21.8
svmp	6640.782	5320.93	0.875	1881.9	17
svmr	5352.394	4232.993	0.916	1773.9	16
svmtune	5645.026	4445.373	0.905	588.5	15
tree1	6875.631	5176.252	0.842	115.8	5
tree2	6624.739	5237.819	0.856	16.4	3.2
treetune	6320.067	4971.764	0.868	41.1	3.8

which on the other hand were not consistent either. Future work could focus on reducing evaluator bias by combining expert opinions with automated evaluation tools to enhance the objectivity and consistency of the interpretability scale. Surveys analyzing the perceptions of interpretability, and objective testing regarding the correctness of performed interpretations by experts may also be explored to objectify interpretability in future work.

Uncertainties play a significant role in machine learning models and can affect both predictive performance and interpretability, and this study is no exception. These uncertainties can be classified as internal (e.g., model assumptions and data quality) or external (e.g., environmental factors influencing the data). Additionally, uncertainties can be parametric, stemming from the model's assumptions about the data distribution, or non-parametric, arising from the variability in the data itself. In our analysis, the internal uncertainties were managed by employing cross-validation techniques, and parameter tuning, while external uncertainties—such as changes in traffic patterns due to seasonal factors—were accounted for through data preprocessing, feature engineering, and stratified sampling. Future work could explore more sophisticated methods to quantify and mitigate these uncertainties, such as Bayesian techniques or ensemble learning approaches, to further enhance the robustness and interpretability of the models.

The rank scale was only developed for structured data and tested with bike traffic data. Even though the scale can be used for other structured datasets, the association of the scale with predictive performance might be different. We only considered numerical outcome variables with L2 loss. The association between predictive performance and the scale may be affected by using other types of loss functions. The scale weighs all parameters in the resulting model equally, regardless of its magnitude and importance in predictive performance. To adjust this, one could weight parameters by effect sizes or by the parameter's contribution to a reduction of the loss function. For the analysis, we used the typical and most used implementations of the stated machine learning algorithms. However, other implementations and variations exist and may also be used more often in the future. Improvements or changes in some of the algorithms may affect both the predictive performance and interpretability of the here analyzed algorithms. Further research is necessary to extend the scale in a way that meaningful parameter weighting can be incorporated and should explore the following areas to build on our findings:

- The interpretability scoring framework developed in this study is tailored for structured data, such as the traffic data used here. Its application to unstructured data (e.g., images, text, or audio) remains an open question. Unstructured data often requires feature extraction techniques like convolutional layers in neural networks, which introduce additional layers of complexity that may affect interpretability. Future research should validate the adaptability of this interpretability framework to unstructured data domains, examining how model parameters and complexity measures can be redefined to capture interpretability in such settings.
- Expand the analysis to include datasets from other fields, such as healthcare or financial data, to validate its universality, generalizability, and flexibility.
- Adapt the interpretability scale to incorporate unsupervised learning algorithms. This includes exploring clustering techniques like k-means and t-SNE, where t-SNE has demonstrated superior performance in capturing complex data structures and clustering accuracy [40], [41]. Future work should also extend the scale to other unsupervised methods such as density estimation and dimensionality reduction, ensuring that the interpretability framework remains applicable across a broader range of machine learning tasks
- Developing more objective measures of interpretability remains crucial. Future research could explore automated scoring methods based on model transparency metrics, such as feature importance distributions or gradient-based attribution methods.
- Investigate the impact of different types of interpretability methods on model performance across various domains.
- Develop new metrics for evaluating the trade-off between predictive performance and interpretability, considering different applications' specific requirements.
- Explore the integration of hybrid models that combine interpretable and non-interpretable components such as in [42] to optimize both predictive performance and transparency.

Building on the limitations and future research directions, we now discuss the results and offer practical implications of the interpretability scale for applied research and industry applications.

VII. DISCUSSION

Our method for defining and measuring interpretability differs to some extent from that proposed by other authors in the literature. Many focus mostly on explainability methods such as [43] or local interpretability methods [44]. Our definition of interpretability is resembled more closely by in-model and global interpretability in [24]. Fuhermore, we couldn't find examples of scales combining objective with subjective metrics resulting in a unified scale.

While our interpretability ranking involves subjective elements, such as expert evaluations of model transparency, we mitigated this subjectivity by cross-referencing rankings with existing literature on machine learning interpretability. A consensus approach was adopted among multiple evaluators to further ensure consistency, reducing individual biases. A fundamental assumption of objectivity in these measures is that they provide reproducible results across different evaluators and datasets, offering a standardized way to assess model transparency. Also, the association of the interpretability scale and predictive performance validates the definition of the scale, as predictive performance is an objective measure. The subjectivity involved in defining interpretability suggests that a universal standard might not be feasible. Our research could be used as a framework adaptable to various contexts. This aligns with the arguments of Carvalho et al. [24], who emphasize the importance of selecting appropriate interpretability methods based on specific application requirements, acknowledging that a one-size-fits-all approach is impractical. Additionally, the impact of domain knowledge on interpretability is strong and important, as models that are easily interpretable in one field might not be so in another due to differences in data complexity and nature. Our study suggests that an interpretability scale is more appropriate for models with a natural way of interpreting results, such as linear models and decision trees, compared to black-box models like neural networks and random forests, whose internal mechanisms are not easily understood. The term 'black box model' describes systems whose internal decision-making processes are opaque. For model engineers, tuning such models is challenging due to the complexity and unknown effects of many hyperparameters. While not all hyperparameters can be tuned, some must be set a priori, such as the number of hidden layers in a neural network.

Our findings highlight a potential link between interpretability and the bias-variance trade-off. In general, the number of model parameters does not directly correspond to bias or variance in black-box models. For instance, a random forest, which consists of many trees, typically exhibits higher variance for a single tree but lower variance when trees are bagged. This complexity, however, results in lower interpretability due to the difficulty in understanding the model. For interpretable models, we suggest that predictive performance is associated with the interpretability rank. There is also a connection to the bias-variance tradeoff. More parameters often lead to lower bias and higher variance, as simpler models (e.g., linear models) tend to have higher bias but are easier to interpret.

The statistical significance of the associations observed between interpretability, the number of parameters, and predictive performance provides robust support for the research hypotheses. Specifically, the positive association between the interpretability scale and predictive performance for interpretable models ($\beta = 0.004$, $p < 0.001$) 5 highlights the relevance of the score for interpretable models.

For black-box models this is not the case, yielding a significant interaction effect of the difference between the effect of interpretable vs. black-box models ($\beta = 0.006$, $p < 0.001$). The logarithmic number of parameters was significantly associated with predictive performance for interpretable and black-box models. Furthermore, the interaction analysis showed that this effect was significantly stronger for interpretable models compared to black box models. Further highlighting the difference between both classes of machine learning models.

The findings aligns with the conclusions of Doshi-Velez and Kim [21], who emphasize the need for a rigorous science of interpretability in machine learning, but still allowing for the subjective nature of interpretability.

Gilpin et al. [22] also discuss the challenges of interpreting complex models and the importance of model transparency. Our findings are consistent with their observations, reinforcing the necessity of balancing predictive performance and interpretability based on the application's requirements. Tjoa and Guan [23] highlight the importance of explainability in medical AI systems to ensure trust and compliance with regulations. Our study extends this argument to other domains, such as urban planning, where explainability can strongly impact decision-making processes.

Rudin [6] argues that simpler, more interpretable models should be preferred whenever possible to avoid the risks associated with black-box models. This viewpoint supports our recommendation to favor interpretable models in scenarios where understanding the decision-making process is crucial. Rudin's work provides a strong case for the use of intrinsically interpretable models, particularly in contexts where the stakes are high, and the consequences of model errors are severe.

Furthermore, the work by Lipton [5] emphasizes the trade-offs between interpretability and performance, suggesting that while simpler models may be less powerful, they offer advantages in terms of transparency and trust. This balance is critical, as users must often weigh the benefits of higher accuracy against the need for interpretability. Our findings are in line with this sentiment, especially for the class of intrinsically interpretable models. The association between the interpretability scale and predictive performance for this class of models highlights the need for complex yet interpretable models associated with higher predictive performance. This includes using interactions, non-linear relationships, and many predictors [45]. Hence, there is a need to create more such models, that are interpretable and still adequately describe the data in its complexity. The scoring system can be used to assess the applicability of machine learning algorithms for a specific problem as a primary prediction model taking both interpretability and predictive performance of the used model into account. The other approach is to use a black-box algorithm and choose the post-hoc interpretation algorithm based on the scoring system. In this setting one interpretation of our approach to achieving model interpretability can be viewed through the lens of feature extraction. By transforming raw data into strongly interpretable features, we can enhance both the interpretability and the predictive performance of the model. This transformation process involves selecting features that have clear semantic meaning or direct relevance to the problem domain, making the model's outputs more understandable. In this way, interpretable feature extraction serves as an important step in explaining the predictive performance of the backend machine learning models, thereby reducing the 'black-box' effect while maintaining performance.

For practitioners, the findings suggest a clear path forward when selecting machine learning models. In applications where transparency is critical—such as healthcare diagnostics or financial risk assessments—practitioners should prioritize models that balance interpretability with sufficient predictive performance. For example, in healthcare, interpretable models like decision trees allow physicians to understand and explain the reasoning behind predictions, fostering trust in AI-driven decisions. Our results suggest that the class of interpretable boosting algorithms, especially those including interactions, provide relatively high predictive performance, often even higher than black-box models. They yield complex, yet interpretable models, making them an excellent choice for traffic management, environmental applications [46], or other domains. However, statistical or mathematical training might be required for practitioners so that the interpretations are accurate. Our results also highlight the importance of model tuning in machine learning which is also discussed a lot in the literature [47], by using sampling methods such as cross-validation or bootstrapping [48], [49].

Shaygan et al. [50] investigate the role of explainable AI in traffic prediction and management, finding that explainability enhances the trust and usability of AI systems in urban planning. Our study's focus on traffic prediction aligns with their findings, demonstrating the practical implications of our results in real-world applications. Zhou et al. underscore the necessity of explainable AI to gain public and stakeholder trust in AI-driven urban infrastructure projects. Models that are easily interpretable in one field might not be so in another due to differences in the complexity and nature of the data. This underscores the need for domain-specific interpretability studies. Although this study focuses on bicycle traffic data as a practical example, the proposed interpretability framework is designed to be adaptable to a wide range of application scenarios. For example, domains such as healthcare, finance, and other sectors that require high interpretability in decision-making models can benefit from our interpretability scale. However, we recognize that the specific models and interpretability methods may need to be adjusted according to the nature of the data, the domain's requirements, and the state of knowledge.

In many machine learning applications, it is often unknown a priori which algorithms and hyperparameter settings will

yield optimal performance for a given dataset. As a result, hyperparameter tuning is typically performed using out-of-sample data, a common practice that involves splitting the data into training and validation sets. However, this method is constrained by the number of available observations, as a significant portion of the dataset must be reserved for validation. The interpretability scale we propose offers a complementary approach: by leveraging the observed association between interpretability and predictive performance, one can approximate a model's performance based on its interpretability score. This allows researchers to narrow down the range of hyperparameters to those most likely to produce models with the desired balance of complexity and predictive accuracy. For example, in regression trees, this may involve adjusting the depth of the tree, while in boosting algorithms, it could guide the number of parameters updated in each iteration. After initial fitting, the interpretability score can be computed, and parameters adjusted accordingly, reducing the need for extensive out-of-sample testing.

ADAPTABILITY TO UNSTRUCTURED DATA

Extending the interpretability scoring system to unstructured data, such as images, text, or audio, presents unique challenges and opportunities. While the scale can be used exactly in the same way, by first assessing the model type and then the number of parameters, unstructured data often requires complex preprocessing. Before predictive modeling, this could include feature extraction techniques, such as convolutional neural networks (CNNs) for images or transformer-based models for text. These preprocessing steps inherently add layers of complexity that may impact interpretability. The scoring system suitable for structured data may not reflect these complexities. To adapt the framework to unstructured data, several modifications can be considered:

- Feature Extraction Transparency: Evaluate the transparency of feature extraction methods. For example, convolutional filters in CNNs could be assessed based on their ability to produce interpretable intermediate outputs, such as activation maps or saliency maps, which highlight important regions of the input data. The scoring system could then be applied to the function transforming the interpretable intermediate output to the final output and not to the full prediction method.
- Redefining Model Complexity: Incorporate metrics that account for the architectural depth of models (e.g., the number of layers in a CNN) or the dimensionality of feature spaces (e.g., embeddings in transformer-based models). These measures could replace or complement parameter counts in the interpretability scale.
- Surrogate Models for Post-hoc Analysis: Apply the scoring system to interpretable surrogate models, such as decision trees or linear regression, to approximate the behavior of complex unstructured data models. These surrogates can provide insight into how features derived from unstructured data contribute to predictions. In this setting, the scoring system can help select and assess appropriate surrogate models.
- Domain-Specific Interpretability: Tailor interpretability evaluations to the specific application. For instance, in medical imaging, interpretability could emphasize regions of interest (e.g., lesions), while in text analysis, attention weights or extracted keywords could serve as proxies for interpretability.

The last point also applies to interpretable structured data. All mentioned modifications can also be applied to non-interpretable or hard-to-interpret structured data. While some proposals are straightforward, others are preliminary and suggest a pathway for adapting the interpretability scale to unstructured data. Future studies should empirically validate these approaches and refine the scoring system to accommodate the unique demands of unstructured data domains.

VIII. CONCLUSION

In summary, our study demonstrates that the interpretability scale can serve as a valuable tool for balancing model complexity with predictive accuracy, providing a practical approach for guiding hyperparameter optimization in machine learning models. By highlighting the relationship between interpretability and performance, we offer a framework that supports the development of models that are not only powerful but also transparent and easier to understand.

Overall, our empirical analysis confirms the importance of understanding the relationship between predictive performance and interpretability in machine learning. The comprehensive analysis of interpretable and non-interpretable machine learning algorithms on a real-world dataset emphasizes the need for explainability in many applications. Our results indicate a negative correlation between the rank of model interpretability and predictive performance for interpretable models, supporting the hypothesis of a trade-off between these two aspects.

However, our empirical examinations also reveal that this relationship does not hold for non-interpretable machine learning algorithms. This finding challenges the generalization of predictive performance versus interpretability trade-off and suggests that for non-interpretable machine learning tools, a focus on predictive performance might be sufficient. This nuanced understanding is critical as it underscores the complexity and contextual dependency of model performance and interpretability, aligning with the views of scholars like Rudin [6] and Lipton [5], who advocate for the careful consideration of model complexity and the associated risks.

In conclusion, our study contributes to the ongoing discourse on the balance between predictive performance and interpretability in machine learning. By providing empirical evidence and practical insights, we hope to guide researchers and practitioners in developing more transparent and effective

TABLE 8. Standard deviation of the models.

Model type sd	Rmse sd	Mae sd	R^2 sd	Parameters sd	Rank sd
gbm1	2960.352	2336.134	0.041	0	0.437
gbm2	3006.476	2374.134	0.049	0	0
gbmtune	16057.71	14602.35	0.04	43.724	0.437
knn1	3532.458	2855.776	0.058	2.348	0
knn2	4301.182	3402.313	0.079	2.348	0
knntune	3334.073	2679.296	0.054	5.636	0
lm	3594.342	2820.662	0.049	0.874	0
mb1	3228.141	2495.873	0.047	20.113	0.437
mb2	3338.574	2639.416	0.041	19.983	0.437
mbsp1	3244.125	2533.214	0.044	1670.924	0.707
mbsp2	3324.086	2609.717	0.039	1098.721	0.437
mbsptune	3255.38	2539.923	0.042	1530.73	0.437
mbtune	3228.169	2495.344	0.047	20.113	0
nn1	5702.573	4736.203	0.054	4.372	0.514
nn2	11274.2	8658.36	0.082	4.372	0.514
nntune	2162.541	1773.401	0.053	10.333	0.961
rf1	3234.844	2564.452	0.04	128109	0.437
rf2	3133.882	2451.617	0.044	79365.35	0
rftune	3143.949	2445.915	0.039	346556.2	0.437
svmp	4234.238	3407.66	0.038	131.273	0
svmr	3344.05	2670.696	0.043	154.069	0
svmtune	3495.519	2760.671	0.051	50.845	0
tree1	3491.157	2679.064	0.085	27.709	0
tree2	3415.43	2717.765	0.058	2.76	0.437
treetune	3448.084	2704.755	0.065	19.243	0.437

TABLE 9. Median of the model types.

Model type sd	Rmse sd	Mae sd	R^2 sd	Parameters sd	Rank sd
gbm1	4548.95	3562.329	0.924	400	18
gbm2	4427.58	3433.531	0.925	800	20
gbmtune	6798.671	5435.556	0.868	400	19
knn1	5227.01	3945.459	0.901	27	6
knn2	6523.55	5026.131	0.852	27	7
knntune	4955.423	3801.839	0.909	73	8
lm	5275.599	4321.563	0.901	11	2
mb1	4682.456	3758.686	0.915	120	9
mb2	5111.355	4092.191	0.913	120	10
mbsp1	4482.593	3540.097	0.919	5817	13
mbsp2	4853.809	3909.323	0.918	4101	12
mbsptune	4500.292	3567.517	0.919	4694	14
mbtune	4682.456	3758.686	0.915	120	11
nn1	4473.366	3513.604	0.911	61	24
nn2	4471.622	3603.331	0.895	61	25
nntune	3606.71	2800.58	0.922	51	26
rf1	5230.276	4108.714	0.915	345304	21
rf2	4928.558	3756.81	0.922	918666	23
rftune	4941.129	3838.327	0.921	487920	22
svmp	5602.092	4526.479	0.884	1836	17
svmr	4474.543	3482.932	0.927	1704	16
svmtune	4862.867	3611.393	0.913	583	15
tree1	5979.658	4088.918	0.876	117	5
tree2	6115.507	4777.199	0.862	15	3
treetune	5059.876	3753.776	0.888	49	4

machine learning models. Further research should investigate the potential for generalising our findings across different domains and data types. Additionally, new methods for enhancing the interpretability of complex models should be explored, and the trade-offs between model accuracy and transparency should be futher refined. Exploring the integration of hybrid models that combine interpretable and non-interpretable components could offer a promising avenue for optimizing both predictive performance and interpretability. We recommend that researchers should prioritize more interpretable models when predictive performance is comparable, and this scale provides a practical tool to implement this interpretability heuristic efficiently and transparently.

TABLE 10. Standard deviation of the model types.

Model type sd	Rmse sd	Mae sd	R^2 sd	Parameters sd	Rank sd
gbm	10250,37	9273,528	0,049	197,831	0,825
knn	3768,624	2986,656	0,072	23,178	0,825
lm	3594,342	2820,662	0,049	0,874	0
mb	3200,109	2494,274	0,044	19,664	0,825
mbsp	3209,118	2510,207	0,041	1477,563	0,825
nn	7941,127	6189,823	0,067	7,389	0,803
rf	3107,812	2438,706	0,04	345452,6	0,825
svm	3679,108	2943,345	0,047	603,39	0,825
tree	3389,731	2648,504	0,07	46,805	0,825

TABLE 11. Median of the model types.

Model type sd	Rmse sd	Mae sd	R^2 sd	Parameters sd	Rank sd
gbm	21068180	3698,344	0,908	400	19
knn	32012570	4116,377	0,889	32	7
lm	27831942	4321,563	0,901	11	2
mb	21925390	3758,686	0,915	120	10
mbsp	20252630	3567,517	0,919	4693	13
nn	16486004	3164,449	0,908	61	25
rf	24414759	3838,327	0,921	48792	22
svm	23647475	3611,393	0,911	1704	16
tree	33907342	4088,918	0,876	49	4

ADDITIONAL TABLES

See Tables 8–11.

ACKNOWLEDGMENT

All statements expressed in this article are the authors' and do not reflect the official opinions or policies of they host affiliations or any of the supporting institutions.

REFERENCES

[1] IDC, Needham, MA, USA. (2024). *IDC's Worldwide AI and Generative AI Spending—Industry Outlook | IDC Blog*. [Online]. Available: https://blogs.idc.com/2024/08/21/idcs-worldwide-ai-and-generative-ai-spending-industry-outlook/

[2] Kinsey. (2024). *The State of AI in Early 2024 | McKinsey*. [Online]. Available: https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

[3] Gartner. (2024). *Gartner Top 10 Strategic Technology Trends for 2024*. [Online]. Available: https://www.gartner.com/en/articles/gartner-top-10-strategic-technology-trends-for-2024

[4] IBM. (2024). *Data Suggests Growth in Enterprise Adoption of AI is Due To Widespread Deployment By Early Adopters, But Barriers Keep 40% in the Exploration and Experimentation Phases*. [Online]. Available: https://newsroom.ibm.com/2024-01-10-Data-Suggests-Growth-in-Enterprise-Adoption-of-AI-is-Due-to-Widespread-Deployment-by-Early-Adopters

[5] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery," *Queue*, vol. 16, no. 3, pp. 31–57, Jun. 2018, doi: 10.1145/3236386.3241340.

[6] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Mach. Intell.*, vol. 1, no. 5, pp. 206–215, May 2019. [Online]. Available: https://www.nature.com/articles/s42256-019-0048-x

[7] N. Jo, S. Aghaei, J. A. Benson, A. Gómez, and P. Vayanos, "Learning optimal fair decision trees: Trade-offs between interpretability, fairness, and accuracy," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, New York, NY, USA, Aug. 2023, pp. 181–192, doi: 10.1145/3600211.3604664.

[8] E. Mariotti, J. M. Alonso Moral, and A. Gatt, "Exploring the balance between interpretability and performance with carefully designed constrainable neural additive models," *Inf. Fusion*, vol. 99, Nov. 2023, Art. no. 101882. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253523001987

[9] J. Wanner, L.-V. Herm, K. Heinrich, and C. Janiesch, "Stop ordering machine learning algorithms by their explainability! An empirical investigation of the tradeoff between performance and explainability," in *Responsible AI and Analytics for an Ethical and Inclusive Digitized Society* (Lecture Notes in Computer Science), vol. 12896, D. Dennehy, A. Griva, N. Pouloudi, Y. K. Dwivedi, I. Pappas, and M. Mäntymäki, Eds., Cham, Switzerland: Springer, 2021, pp. 245–258. [Online]. Available: https://link.springer.com/10.1007/978-3-030-85447-822

[10] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," 2019, *arXiv:1910.10045*.

[11] G. R. Sridhar, A. V. S. Prasad, and G. Lakshmi, "Scope and caveats: Artificial intelligence in gastroenterology," *Artif. Intell. Gastroenterol.*, vol. 25, no. 14, pp. 1666–1683, Apr. 2019.

[12] D. Gunning and D. Aha, "DARPA's explainable artificial intelligence," *AI Mag.*, vol. 40, no. 2, pp. 44–58, Jun. 2019. [Online]. Available: https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2850

[13] H. K. Dam, T. Tran, and A. Ghose, "Explainable software analytics," in *Proc. IEEE/ACM 40th Int. Conf. Softw. Engineering: New Ideas Emerg. Technol. Results (ICSE-NIER)*, New York, NY, USA, May 2018, pp. 53–56, doi: 10.1145/3183399.3183424.

[14] F. Jaotombo, L. Adorni, B. Ghattas, and L. Boyer, "Finding the best trade-off between performance and interpretability in predicting hospital length of stay using structured and unstructured data," *PLoS ONE*, vol. 18, no. 11, Nov. 2023, Art. no. e0289795. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0289795

[15] J.-P. Kucklick and O. Müller, "Tackling the accuracy-interpretability trade-off: Interpretable deep learning models for satellite image-based real estate appraisal," *ACM Trans. Manage. Inf. Syst.*, vol. 14, no. 1, pp. 1–24, Jan. 2023, doi: 10.1145/3567430.

[16] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, Jan. 2021. [Online]. Available: https://www.mdpi.com/1099-4300/23/1/18

[17] R. Confalonieri, L. Coba, B. Wagner, and T. R. Besold, "A historical perspective of explainable artificial intelligence," *WIREs Data Mining Knowl. Discovery*, vol. 11, no. 1, p. e1391, Jan. 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1391

[18] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, Aug. 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.

[19] S. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates Inc., Jan. 2017, pp. 4768–4777.

[20] R. Rodríguez-Pérez and J. Bajorath, "Interpretation of machine learning models using Shapley values: Application to compound potency and multi-target activity predictions," *J. Computer-Aided Mol. Design*, vol. 34, no. 10, pp. 1013–1026, Oct. 2020, doi: 10.1007/s10822-020-00314-0.

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

[22] 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. [Online]. Available: https://ieeexplore.ieee.org/document/8631448

[23] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 4793–4813, Nov. 2021. [Online]. Available: https://dr.ntu.edu.sg/handle/10356/154295

[24] 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. [Online]. Available: https://www.mdpi.com/2079-9292/8/8/832

[25] P. Ryus, E. Ferguson, K. Laustsen, R. J. Schneider, F. R. Proulx, T. Hull, and L. Miranda-Moreno, "Methods and technologies for pedestrian and bicycle volume data collection," NCHRP web-only document, Tech. Rep. NCHRP Project 07-19, 2014. [Online]. Available: https://trid.trb.org/View/1342013

[26] B. Ripley. (Feb. 2023). *Tree: Classification and Regression Trees*. [Online]. Available: https://CRAN.R-project.org/package=tree

[27] T. Therneau, B. Atkinson, B. Ripley, and B. Atkinson. (Dec. 2023). *Rpart: Recursive Partitioning and Regression Trees*. [Online]. Available: https://CRAN.R-project.org/package=rpart

[28] A. Beygelzimer, S. Kakadet, J. Langford, A. Sunil, D. Mount, and S. Li. (Jan. 2024). *FNN: Fast Nearest Neighbor Search Algorithms and Applications*. [Online]. Available: https://CRAN.R-project.org/package=FNN

[29] K. Schliep, K. Hechenbichler, and A. Lizee. (Mar. 2016). *Kknn: Weighted K-Nearest Neighbors*. [Online]. Available: https://CRAN.R-project.org/package=kknn

[30] T. Hothorn, P. Buehlmann, T. Kneib, M. Schmid, B. Hofner, F. Otto-Sobotka, F. Scheipl, and A. Mayr. (Apr. 2024). *Mboost: Model-Based Boosting*. [Online]. Available: https://CRAN.R-project.org/package=mboost

[31] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, F. Leisch, and C.-C. Chang. (Dec. 2023). *E1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. [Online]. Available: https://CRAN.R-project.org/package=e1071

[32] A. Karatzoglou, A. Smola, K. Hornik, N. I. Australia (NICTA), M. A. Maniscalco, and C. H. Teo. (Jan. 2023). *Kernlab: Kernel-Based Machine Learning Lab*. [Online]. Available: https://CRAN.R-project.org/package=kernlab

[33] G. Ridgeway, D. Edwards, B. Kriegler, S. Schroedl, H. Southworth, B. Greenwell, B. Boehmke, and J. Cunningham, (Jun. 2024). *Gbm: Generalized Boosted Regression Models*. [Online]. Available: https://CRAN.R-project.org/package=gbm

[34] A. Cutler, M. Wiener, L. Breiman, and A. Liaw. (May 2022). *RandomForest: Breiman and Cutler's Random Forests for Classification and Regression*. [Online]. Available: https://CRAN.R-project.org/package=randomForest

[35] S. Fritsch, F. Guenther, M. N. Wright, M. Suling, and S. M. Mueller. (Feb. 2019). *Neuralnet: Training of Neural Networks*. [Online]. Available: https://CRAN.R-project.org/package=neuralnet

[36] B. Ripley and W. Venables. (May 2023). *Nnet: Feed-Forward Neural Networks and Multinomial Log-Linear Models*. [Online]. Available: https://CRAN.R-project.org/package=nnet

[37] S. Wood. (Dec. 2023). *Mgcv: Mixed GAM Computation Vehicle With Automatic Smoothness Estimation*. [Online]. Available: https://CRAN.R-project.org/package=mgcv

[38] H. Wickham, W. Chang, L. Henry, T. L. Pedersen, K. Takahashi, C. Wilke, K. Woo, H. Yutani, D. Dunnington, and T. van den Brand. (Apr. 2024). *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. [Online]. Available: https://CRAN.R-project.org/package=ggplot2

[39] H. Wickham, R. Francois, L. Henry, K. Müller, and D. Vaughan. (Nov. 2023). *Dplyr: A Grammar of Data Manipulation*. [Online]. Available: https://CRAN.R-project.org/package=dplyr

[40] K. Dhalmahapatra, R. Shingade, H. Mahajan, A. Verma, and J. Maiti, "Decision support system for safety improvement: An approach using multiple correspondence analysis, t-SNE algorithm and K-means clustering," *Comput. Ind. Eng.*, vol. 128, pp. 277–289, Feb. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360835218306478

[41] H. Kwon, Z. A. Ali, and B. M. Wong, "Harnessing semi-supervised machine learning to automatically predict bioactivities of per- and polyfluoroalkyl substances (PFASs)," *Environ. Sci. Technol. Lett.*, vol. 10, no. 11, pp. 1017–1022, Nov. 2023, doi: 10.1021/acs.estlett.2c00530.
[42] F. Obster, J. Brand, M. Ciolacu, and A. Humpe, "Improving boosted generalized additive models with random forests: A zoo visitor case study for smart tourism," *Proc. Comput. Sci.*, vol. 217, pp. 187–197, Jan. 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S187705092202292X
[43] P. Schmidt and F. Biessmann, "Quantifying interpretability and trust in machine learning systems," 2019, *arXiv:1901.08558*.
[44] D. Slack, S. A. Friedler, C. Scheidegger, and C. D. Roy, "Assessing the local interpretability of machine learning models," 2019, *arXiv:1902.03501*.
[45] F. Obster, C. Heumann, H. Bohle, and P. Pechan, "Using interpretable boosting algorithms for modeling environmental and agricultural data," *Sci. Rep.*, vol. 13, no. 1, p. 12767, Aug. 2023. [Online]. Available: https://www.nature.com/articles/s41598-023-39918-5
[46] F. Obster, H. Bohle, and P. M. Pechan, "The financial well-being of fruit farmers in Chile and Tunisia depends more on social and geographical factors than on climate change," *Commun. Earth Environ.*, vol. 5, no. 1, pp. 1–12, Jan. 2024. [Online]. Available: https://www.nature.com/articles/s43247-023-01128-2
[47] F. Hourdin, T. Mauritsen, A. Gettelman, J.-C. Golaz, V. Balaji, Q. Duan, D. Folini, D. Ji, D. Klocke, Y. Qian, F. Rauser, C. Rio, L. Tomassini, M. Watanabe, and D. Williamson, "The art and science of climate model tuning," *Bull. Amer. Meteorol. Soc.*, vol. 98, no. 3, pp. 589–602, Mar. 2017. [Online]. Available: https://journals.ametsoc.org/view/journals/bams/98/3/bams-d-15-00135.1.xml
[48] J.-H. Kim, "Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap," *Comput. Statist. Data Anal.*, vol. 53, no. 11, pp. 3735–3745, Sep. 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167947309001601
[49] T. Fushiki, "Estimation of prediction error by using K-fold cross-validation," *Statist. Comput.*, vol. 21, no. 2, pp. 137–146, Apr. 2011, doi: 10.1007/s11222-009-9153-8.
[50] M. Shaygan, C. Meese, W. Li, X. G. Zhao, and M. Nejad, "Traffic prediction using artificial intelligence: Review of recent advances and emerging opportunities," *Transp. Res. C, Emerg. Technol.*, vol. 145, Dec. 2022, Art. no. 103921. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0968090X22003345

FABIAN OBSTER received the bachelor's degree in mathematics and statistics from LMU Munich and the master's degree in mathematics in data science from Technical University Munich. He is currently a Research Associate with the Department of Business Administration, University of the Bundeswehr Munich, and the Department of Statistics, Ludwig Maximilians University Munich. His main areas of research interests include machine learning, statistical modeling, and environmental sciences.

MONICA I. CIOLACU received the Engineering degree in electronic engineering, in 1996, and the Ph.D. degree in electronics, telecommunications, and information technology from the University Politehnica of Bucharest, Romania, in 2020.

She is currently a Lecturer and a Researcher with the University of Passau, Germany. Her research focus is learning and teaching with and about generative AI, research methods, and technology-enhanced learning. She has eight years of engineering experience in product marketing and business development from her tenure with Kontron Embedded Modules Germany. Her research interests include education 4.0/5.0, learning analytics, Industry 4.0/5.0, biofeedback, and the development of intelligent blended learning environments. She is an active contributor to the academic community, serving on various IEEE technical conference committees, and acting as a Reviewer for EDUCON, IEEE Access, IEEE Education, TALE, and SIITME.

ANDREAS HUMPE (Member, IEEE) received the four master's degrees in finance and investment management, intelligent systems and robotics, advanced manufacturing systems, and astrophysics, and the Ph.D. degree in econometrics. He is currently a Professor of mathematics and finance with the Department of Tourism, Munich University of Applied Sciences, Germany, and part of the Institute for Applications of Machine Learning and Intelligent Systems (IAMLIS). His main areas of research interests include environmental sciences, transportation, and mathematical modeling.